CABI BIOSCIENCE
EGHAM LIBRARY

Commonwealth Bureau of Plant Breeding and Genetics

TECHNICAL COMMUNICATION No. 15

Field Trials II:
The Analysis of Covariance

by

JOHN WISHART, M.A., D.Sc.
(Reader in Statistics in the University of Cambridge)

School of Agriculture
Cambridge
England

FIELD TRIALS II:

THE ANALYSIS OF COVARIANCE

Introduction

The earlier bulletin, *Field Trials: their Layout and Statistical Analysis*, published in 1940 (Imperial (now Commonwealth) Bureau of Plant Breeding and Genetics) was intended as an elementary exposition of the standard methods that normally form the content of a first course of instruction for the agricultural student in this subject. There is nothing in that bulletin which has been superseded by later developments, and it should still be consulted by the beginner. There are directions, however, in which it is desirable to go in order to carry the instruction of the plant breeder or agricultural statistician to a more advanced level. More than one supplementary bulletin is likely to be called for; for the present the most urgent task appears to be to give an exposition of the methods commonly grouped under the title "Analysis of Covariance". To quote from the earlier bulletin (p. 35): "There is the question of taking into simultaneous consideration two or more observational variables from the same plot with a view, either to the elaboration of methods for taking account of soil fertility variations in a more complete fashion than is possible merely by eliminating block, or row and column, differences, or to the further elucidation of the nature of the facts sought to be learnt from the experiment. This brings in the calculations involved in the analysis of co-variance procedure." To study this question is the object of the present bulletin. The methods used will be such as can easily be superimposed on the ordinary structure of the simple experiment; for example, this may be laid out in randomized blocks or in a simple Latin square, the sole difference being that multiple measurements on the plots (which are commonly made anyhow by the plant breeder) are taken into account in order to refine the analysis, or to learn more concerning varietal or treatment differences. Throughout, the mention in brackets of 1940, with a page number, will be a reference to the bulletin whose title is given in full at the beginning of this section.

Measurement of Regression

Let us start by assuming that two distinct measurements have been made on each of a number of equal-sized plots which have been uniformly treated in a yield trial; that is, we are to imagine the plots laid out over an area uniform in its fertility, and on which there are no imposed varietal or treatment differences. These measurements might be the number of tillers in a wheat crop in the spring, when the plants may be said to be established, and the final yield of grain at harvest; or they might be the number of plants at harvest in a crop of sugar beet and the yield of beet, or of sugar. The two characters measured will be referred to as the *variates*. Following the former practice (1940, p. 4) we shall denote the measures of the *x-variate* by x_1, x_2, x_3, etc. and those of the *y-variate* by y_1, y_2, y_3, etc. Now we know that if we add up the values of x available and divide by their number we get the *mean*, denoted by $\bar{x}$, a quantity which may be taken as an estimate of the true value of the x-variate. Further, we know that if we perform a calculation which is equivalent to subtracting this mean from each of the values of x_1, x_2, x_3, etc., squaring these deviations, adding up and dividing by one less than the number of observations (i.e. by the *degrees of freedom* D.F.) we get an estimate, denoted by s^2, of the *variance*, its square root being an estimate of the *standard deviation*. Practically, the way to proceed is to use Method II, or in some cases Method III (1940, p. 5).

Now the x and y variates may or may not be related. In one of the cases illustrated above it may happen that the yield of wheat grains is large when the number of tillers is large, or small when the number is small. Furthermore it is reasonable in this case to consider the number of tillers to be one, at least, of the factors determining the final yield. In such cases we may speak

of the y-variate as being dependent on the value of the x-variate, and we may call y the *dependent* variate. The same considerations may apply in the second of our illustrations. The x-variate in such cases is often called, by contrast, the *independent* variate, but this terminology is not very satisfactory because we have to distinguish between cases where y and x are related and cases where no relation exists, i.e. where y and x are *independent* of one another. It may be better to speak of y as the *affected* variate and x as the *affecting* variate. Now another case can arise of two variates being related where it is not possible to say which of the two is the affecting variate. Suppose we measure the yield of both grain and straw in a wheat experiment. We shall find that they are related in the sense of the present paragraph, but who is to say that it is the amount of straw that determines the amount of grain, or the reverse? The methods of statistics do serve in some measure to distinguish the different cases, but they are both based on the same kind of fundamental calculation.

To investigate the nature of the relationship let us bring together in adjacent columns the x-deviations and the y-deviations, so that $x_1 - \bar{x}$ will be alongside $y_1 - \bar{y}$, and $x_2 - \bar{x}$ alongside $y_2 - \bar{y}$, etc. In the following simple example x is the number of bean shoots in March and y the yield of beans at harvest (in hectograms) in a Cambridge experiment:—

TABLE I

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})\ (y-\bar{y})$
	173	10·7	24	1·1	26·4
	148	9·8	— 1	0·2	— 0·2
	153	9·5	4	— 0·1	— 0·4
	122	8·4	—27	— 1·2	32·4
Total	596	38·4			Total 58·2 (÷ 3)
$\bar{x}$	149	$\bar{y}$ 9·6			M.P. = 19·4

In the last column we have multiplied together the corresponding deviations and added the products, giving a total of 58·2. By analogy with the process of summing the squares of deviations to produce a Sum of Squares (denoted, for short, by S.S. and particularized as (x^2) or (y^2)), we shall call this result a Sum of Products (denoted by S.P. and particularised as (xy)). There is one important difference which should be noted. Individual squares of deviations must be positive always, and their sum a positive quantity. On the other hand individual products of deviations may be positive or negative, and likewise their sum may have either sign. If there is a *positive* relationship between x and y in the sense that the measures tend to go up or down together the individual products will tend to be positive, and their total will certainly be a positive quantity; but if the reverse is the case, for example, if on the whole large values of y are associated with small values of x, and *vice versa*, individual products will tend to be negative, and their total will certainly be a negative quantity. A third case may be noted; if the variates are not associated it will be a matter of chance whether a positive x-deviation is paralleled by a positive or negative y-deviation. The individual products will then occur with a mixture of positive and negative signs, and their total will be small. We have learnt to distinguish between the *population* of measurements which would be generated if we went on taking observations of the same character indefinitely under the same conditions, and the actual *sample* of measures we have obtained. In the present case we should find that the average of the products of the deviations would in the long run be zero if there was a complete absence of association between the x and y variates. In any sample as ordinarily observed the sum of products (as now defined) will only rarely be exactly zero; rather will it tend to fluctuate about the zero point (in the absence of a relationship between x and y), being sometimes positive and sometimes negative.

In our example the S.P. is + 58·2. Now just as we divide a S.S. by the number of D.F. to obtain a *mean square* (M.S.), so here we can divide by the D.F. and get a *mean product* (M.P.). This M.P. is the sample estimate of the *covariance*, and is the only new quantity we need to calculate when two variates are considered together. It now becomes important to be able to distinguish between the case where the M.P. is small because of absence of association between x and y, and the case where it is large enough to lead us to assert that a significant association

probably exists. Furthermore, we find it useful to calculate coefficients to measure the strength of the association. One such quantity is the estimated *regression coefficient*, generally denoted by b, and derived as (xy) divided by the S.S. of the affecting variate. We shall speak of b as the estimated regression coefficient of y on x, since we are using y for the affected variate and x for the affecting variate.

(x^2) for the data in Table 1 is 1322, and so we have $b = 58{\cdot}2/1322 = 0{\cdot}044$. Its meaning will be made clear if we write down the *regression equation*

$$Y = 9{\cdot}6 + 0{\cdot}044\,(x - 149).$$

The mean of the y-series is 9·6 and that of the x-series 149. Assuming a uniform (or linear) rate of change, we may expect that for every unit increase or decrease of x the corresponding value of y will be increased or decreased by 0·044. b is a *rate of change*, and is measured in the units of the data; in the present case it will represent a change of 0·044 hectograms of yield per shoot. Take the first line of Table 1, where x is 173. Substituting in the above equation we have:—

$$\begin{aligned} Y &= 9{\cdot}6 + 0{\cdot}044\,(173 - 149) \\ &= 9{\cdot}6 + 0{\cdot}044\,(24) \\ &= 9{\cdot}6 + 1{\cdot}06 = 10{\cdot}66. \end{aligned}$$

Note that this is a calculated value, which is why we have used the symbol Y instead of y. The corresponding actual value of y is 10·7. The reader should verify that the remaining values of Y are as set out in Table 2. The total of the values of Y should be equal to 38·4 exactly, as with the values of y; the discrepancy of 0·01 is due to retaining only two decimal places in Y; to go further is not worth while, particularly as b is not exactly 0·044.

TABLE 2

	y	Y	$y - Y$	$(y - Y)^2$
	10·7	10·66	0·04	0·0016
	9·8	9·56	0·24	0·0576
	9·5	9·78	− 0·28	0·0784
	8·4	8·41	− 0·01	0·0001
Total	38·4	38·41	− 0·01	0·1377

TEST OF SIGNIFICANCE

In Table 2 we have proceeded to subtract each Y from its corresponding y, then to square the differences and add them up, getting the total 0·1377. Being made up of the deviations $y - Y$, or *residuals*, this total may be called the deviation sum of squares (d^2). It differs from the total of the squares of $y - \bar{y}$ by the use of both $\bar{y}$ and b in its calculation. In any case it will always be smaller. We may construct a mean square out of it by dividing by the number of D.F., which will in this case be *two less* than the number of squares that have been summed. This yields in our example an estimate of the residual variance, equal to 0·1377 ÷ 2, or 0·0688. The square root, namely 0·26, is often used as the *standard error* appropriate to the regression equation. But it is not the standard error of b; to estimate this quantity we have to divide the M.S. by (x^2), i.e. by 1322. This yields 0·000052, of which the square root is 0·0072. The object of determining a standard error for b is to be able to say whether it is significant or not, or to compare one b with another. In the former case we should note that a sample will normally yield a value of b different from zero even if x and y are entirely unrelated in the population from which the measures have come. The actual value will, however, be small, and we can measure its strength by dividing it by its standard error. This gives in our example the number 6·1, i.e. b is 6·1 times its standard error. The test of significance is of exactly the same character as for a sample mean (1940, p. 6). This number 6·1 is t, and inspection of the table of the t-distribution with D.F. = 2 (Fisher and Yates, *Statistical Tables*, Table III) shows that a value of 4·3 would be required for significance at the 5 *per cent.* point. Our example, therefore, has demonstrated a significant positive association between x and y.

The arithmetic of this example has been gone through in detail in order to make the procedure clear, both as to the calculation of the regression coefficient and the testing of its significance. But we may now go back and show how the calculations may be most expeditiously carried out. Just as (x^2) is in practice calculated by summing the squares of the x's and subtracting a *correction factor* (C.F.), namely the square of the total of the x's divided by the number of observations (1940, p. 5, Method II), so (xy) may be calculated from the sum of the products of x and y. The correction factor here is the product of the total of the x's and the total of the y's, divided by the number of observations. The calculations are shown in Table 3.

TABLE 3

x	y	xy
173	10·7	1851·1
148	9·8	1450·4
153	9·5	1453·5
122	8·4	1024·8
Total 596	38·4	5779·8
		− 5721·6
		58·2

$$\text{Correction factor} = \frac{596 \times 38{\cdot}4}{4} = 5721{\cdot}6$$

We may simplify the calculations still further according to Method III (1940, p. 5). The reader may verify that the same result will ensue from dropping the 1 in the hundreds place of the x-series, i.e. by subtracting 100 from each entry. We could equally subtract a common number from the y-series, or do both in combination. The logical end to this process is to subtract the actual means as in Table 1. Given a calculating machine it is generally quickest to use the numbers as they stand, as in Table 3. The further stage is to calculate (d^2). This is obtained from (y^2) by subtracting an additional quantity, easily remembered in the following way. To calculate b we work out $(xy)/(x^2)$. Then from (y^2) we subtract $(xy)^2/(x^2)$, i.e. the numerator of b has to be squared before dividing by (x^2). The same result may be obtained by multiplying b by (xy) provided we have calculated b to a sufficient number of decimal places, since (xy) may be a large number and we want the result to the same accuracy as (y^2). Alternatively we may multiply b^2 by (x^2). In our example $(xy) = 58{\cdot}2$ and $(x^2) = 1322$. We then have $58{\cdot}2^2/1322 = 2{\cdot}5622$, and (y^2) by the usual methods comes to 2·7 (exact). Then $(d^2) = 2{\cdot}7 - 2{\cdot}5622 = 0{\cdot}1378$, compared with 0·1377 in Table 2, which was not, as already noted, an exact result. The basic quantities to calculate, then, are (x^2), (y^2) and (xy), the last being the only new one. The calculations of this paragraph are all that are required unless for some special reason the individual values of Y are needed. For short we shall denote $(xy)^2/(x^2)$ by (b^2), this being the sum of squares "due to" regression.

The entire calculations on the data provided by the first two columns of Table 1 are given below:—

TABLE 4

$T(x) = 596 \quad \bar{x} = 149 \qquad$ C.F. $(x) = 596^2/4 = 88{,}804$
$T(y) = 38{\cdot}4 \quad \bar{y} = 9{\cdot}6 \qquad$ C.F. $(y) = 38{\cdot}4^2/4 = 368{\cdot}64$
C.F. $(xy) = 596 \times 38{\cdot}4/4 = 5{,}721{\cdot}6$
$(x^2) = 90{,}126 - 88{,}804 = 1{,}322$
$(y^2) = 371{\cdot}34 - 368{\cdot}64 = 2{\cdot}7$
$(xy) = 5{,}779{\cdot}8 - 5{,}721{\cdot}6 = 58{\cdot}2$
$b = 58{\cdot}2/1322 = 0{\cdot}044 \quad (b^2) = 58{\cdot}2^2/1322 = 2{\cdot}5622$
$(d^2) = 2{\cdot}7 - 2{\cdot}5622 = 0{\cdot}1378$
Divide by D.F. (2) 0·0689 Square root = 0·26
Divide by (x^2) 0·000052 Square root = 0·0072
$t = 0{\cdot}044/0{\cdot}0072 = 6{\cdot}1^*$, D.F. = 2, $P < {\cdot}05$
$Y = 9{\cdot}6 + 0{\cdot}044\,(x - 149)$ with S.E. 0·26.

It is instructive to set this out in the alternative form of an *analysis of variance* of the y-variate, because this is the form we shall be using when we come to the adjusting methods used in

covariance problems. What we do is to separate off from (y^2) a part "due to" the regression (this is the above (b^2) having 1 D.F.), leaving a remainder representing "deviations from" the regression, or residual (this is the above (d^2) having D.F. = number of observations − 2). The construction of mean squares, and their testing by the *variance ratio test*, follow in the ordinary way (1940, p. 12). The table is as follows:—

TABLE 5.—ANALYSIS OF VARIANCE

Variation	D.F.	S.S.	M.S.	V.R.	b
Regression	1	2·5622	2·5622	37·2*	0·044
Residual	2	0·1378	0·0689		
Total	3	2·7000			

V.R. denotes variance ratio, and is calculated as the ratio of the two numbers in the preceding column. The D.F. are 1 and 2 respectively, and the star denotes that this value is significant at the 5 *per cent.* point (Fisher and Yates, Table V). Since D.F. = 1 for the numerator, the V.R. is just the square of the value of t already found, namely 6·1, and thus the two tests are equivalent. But Table 5 involves the smallest amount of calculation, the necessary steps being shown in the relevant lines of Table 4.

MEASUREMENT OF CORRELATION

To return now to the case where it is impossible to say which is the affected and which the affecting variable, it should be clear that there is nothing in the procedure of the previous section to prevent our interchanging the roles of x and y. We have obtained an estimated regression coefficient which can be denoted by $b_{y \cdot x}$ to denote that it is for "y on x". A similar coefficient $b_{x \cdot y}$ for "x on y" could be calculated in the same way, merely reversing x and y in the formulæ. But neither can be said in the present case to have much meaning in the sense in which we like to interpret a regression coefficient, namely a rate of change in the affected variable per unit change in the affecting variable. A symmetrical coefficient of association may be constructed by taking a suitable mean of $b_{y \cdot x}$ and $b_{x \cdot y}$. If we take their geometric mean, i.e. multiply them together and extract the square root, giving it the sign of (xy), we get the estimated *correlation coefficient*, generally denoted by r, to which the subscript xy may be attached when it is desired to differentiate it from any other correlation coefficient, as when a third variable z might be present. Given (x^2), (y^2) and (xy), which are needed anyhow, the direct way to calculate r is to divide (xy) by the geometric mean of (x^2) and (y^2). The resulting coefficient is a pure number, and is given the sign of (xy). It may be shown to have a maximum value of + 1 for perfect positive association between x and y, and a minimum value of − 1 for perfect negative association. For complete absence of association the value will be zero, since it is determined from (xy). This would, however, relate to the entire population from which our sample is taken. As with b, a sample, even in such a case, will yield a small value of (xy) from which will result a value, positive or negative, of r. We thus require a test of significance before we can say that the calculations point to the existence of a real relationship between x and y. There are two ways in which this can be done. The first is to calculate r and look it up directly in Fisher and Yates, Table VI, for D.F. equal to the number of observations less 2. In the case of our example we have $\sqrt{(1322 \times 2{\cdot}7)} = 59{\cdot}7$ and therefore $r = 58{\cdot}2/59{\cdot}7 = 0{\cdot}97$, with D.F. = 2. Table VI gives 0·95 for the 5 *per cent.* point, and r is therefore significant at this level. The other way is to construct an analysis of variance, in which (y^2) is divided into the fractions $r^2\,(y^2)$ and $(1 - r^2)\,(y^2)$, with D.F. 1 and 2 respectively, and compare the M.S.'s by the variance ratio. But from the definition we see that $r^2\,(y^2)$ is the same thing as $(xy)^2/(x^2)$, which is our (b^2) of Table 4. Thus the analysis of variance table is exactly the same as Table 5, and we see that, so long as we are testing the significance of departure from a real zero value, the tests of $b_{y \cdot x}$ and r are equivalent. Furthermore, we can see that, with Table 5 constructed, r can be calculated from the S.S. alone, for $r^2 = 2{\cdot}5622/2{\cdot}7 = 0.949$, whence $r = 0{\cdot}97$. Equally, of course, we could have constructed an analysis of variance in which (x^2) was divided into the fractions $r^2\,(x^2)$ and $(1 - r^2)\,(x^2)$. This would be equivalent to a regression analysis of variance in which x was taken as the affected variate and y as the affecting variate.

Difference Between Two Regression Coefficients from Independent Samples

As we are only concerned with the methods needed for the analysis of covariance procedure, we shall not touch on the rather more complex situation which arises out of the last two sections, in the case where the true value of the regression or correlation coefficient is not zero. But there is a standard test which is important for our present purpose; this arises where two samples which may be taken to be independent yield separate regression estimates, and we desire to know whether the coefficients differ significantly from one another. To examine this question we do not need to know the true values of the regression coefficients, but merely assume them to be equal, and we then calculate the chance of a difference in estimated regression coefficients, under the assumed condition, being as great as, or greater than, that actually observed. One further assumption is usually made, and that is, that whatever may be the difference of the coefficients, the residual variances are the same in the two populations. So long as this condition holds, at any rate approximately, we are correct in ascribing a significant result to the difference between the coefficients; otherwise the issue would be in doubt as to the respects in which the two populations differed. The situation is the same as when we are testing for the difference between the means of two samples by the extended t-test, and indeed, since we have already shown that the significance test for b can be expressed as a t-test, it follows that the difference between two estimated b's will yield, similarly to the case of two means, an extended t-test. The problem is treated in this way by R. A. Fisher (*Statistical Methods for Research Workers*, § 26.1). It is more appropriate to what follows, however, to present the test now in its analysis of variance form.

From the first sample of n_1 observations we calculate (x_1^2), (x_1y_1) and (y_1^2), from which we deduce (b_1^2) with 1 D.F. and (d_1^2) with $n_1 - 2$ D.F. Similarly from the second sample of n_2 observations a corresponding calculation yields (b_2^2) with 1 D.F. and (d_2^2) with $n_2 - 2$ D.F. Now add together the sums of squares and products for both samples, yielding $(x_1^2) + (x_2^2)$, $(x_1y_1) + (x_2y_2)$ and $(y_1^2) + (y_2^2)$, each having $n_1 + n_2 - 2$ D.F. Treat this as a single sample and deduce by the foregoing methods $(\bar{b}^2)$, with 1 D.F. and $(\bar{d}^2)$ with $n_1 + n_2 - 3$ D.F. The bar over b and d denotes that we are following out the consequences of fitting the average regression for the combined data. The value of $\bar{b}$, the average regression coefficient, is in fact

$$\{(x_1y_1) + (x_2y_2)\}/\{(x_1^2) + (x_2^2)\}.$$

Considering now only the deviations from the average regression, namely $(\bar{d}^2)$, having $n_1 + n_2 - 3$ D.F., the above analysis shows it to be composed essentially of two parts which can be proved to be independent of one another. The first is $(b_1^2) + (b_2^2) - (\bar{b}^2)$, having 1 D.F., which can be shown by a little algebraic manipulation to be equal to $(b_1 - b_2)^2$ multiplied by the factor $(x_1^2)\ (x_2^2)/\{(x_1^2) + (x_2^2)\}$. The other part is $(d_1^2) + (d_2^2)$, having $n_1 + n_2 - 4$ D.F. Divide the two parts by the respective numbers of D.F. (this makes no difference to the first part, which has 1 D.F. only), and then take their ratio. The result is a variance ratio with D.F. 1 and $n_1 + n_2 - 4$, and if significant will demonstrate that b_1 and b_2 are significantly different from one another.

The calculations will be easily followed from a numerical example. For the first sample we shall take the data already illustrated; for the second we use an independent sample of 5 pairs of the same measures which gave the sums of squares and products set out in Table 6:—

Table 6. Comparison of Regressions

Sample	D.F.	(x^2)	(xy)	(y^2)	D.F.	(b^2)	D.F.	(d^2)
1	3	1322	58·2	2·70	1	2·5622	2	0·1378
2	4	1876	60·8	2·44	1	1·9705	3	0·4695
					2	4·5327	5	0·6073
1 + 2	7	3198	119·0	5·14	1	4·4281		
					1	0·1046		

M.S. $(d) = 0{\cdot}6073/5 = 0{\cdot}1215$

V.R. for $b_1 - b_2 = 0{\cdot}1046/0{\cdot}1215 = 0{\cdot}86$. D.F. = 1 and 5.

The difference between the regression coefficients (0·044 and 0·032) is not significant.

Explanation.—As with Sample 1 (see Table 4), so in Sample 2 we have $60{\cdot}8^2/1876 = 1{\cdot}9705$ and $2{\cdot}44 - 1{\cdot}9705 = 0{\cdot}4695$. Now $1322 + 1876 = 3198$ etc. and $119{\cdot}0^2/3198 = 4{\cdot}4281 = (\bar{b}^2)$ ($(\bar{d}^2)$ need not be written down). Finally $4{\cdot}5327 - 4{\cdot}4281 = 0{\cdot}1046$. Note that $(1322)\ (1876)/3198 = 775{\cdot}5$, which when multiplied by 0·0001348, the square of the difference between $b_1 = 0{\cdot}04402$ and $b_2 = 0{\cdot}03241$, gives 0·1045, showing that at least four significant figures are required for the b's for this calculation to check with the easier one of the table. Should we desire the standard error of $b_1 - b_2$ this may be obtained by noting that V.R. $= t^2 = 0{\cdot}8609$, whence $t = 0{\cdot}928$. As this is the ratio of $b_1 - b_2$, i.e. 0·0116, to its standard error, this last quantity is obtained as $0{\cdot}0116/0{\cdot}928 = 0{\cdot}0125$.

A corresponding test for use when the samples cannot be assumed to be independent is developed later in this bulletin.

Adjustment of the Observations in a Field Trial for Values of an Affecting Variate

The simple calculations of Table 6 can easily be extended when there are more than two samples. The last entry in the column (b^2) will then have D.F. equal to one less than the number of samples. Its M.S., when compared with M.S. (d), this last being a pooled measures from *all* samples, will yield a V.R. whose significance will demonstrate that the set of sample regression coefficients differ significantly among themselves. Such a test is frequently worth making with data suspected of heterogeneity in this respect, and if significant differences in regression are found the measure of residual variance used will be that determined from the separately calculated coefficients of regression. When we have the data of a field trial, laid out in the usual manner, e.g. in randomized blocks or a Latin square, or in some more complex form of lay-out, and wish to take account of another plot variate which may be affecting the measure under consideration, we can usually only obtain one (pooled) measure of error, and therefore only one (pooled) measure of *error* regression. In this case the method of the previous section is applied directly, as shown in the section following this one. A number of cases occur in practice where it is desirable to adopt this procedure. To begin with, the need for any randomized design (1940, pp. 7–15) arises from the fact that similar plots over even a uniform area are very variable in their fertility, and if we are to have an accurate comparison of, say, a number of varieties, we must not only see that these are replicated, in order to obtain an estimate of the error of the comparisons which we shall make of the variety means, but also we must endeavour, by choice of a suitable design, to keep the error down as much as possible. To illustrate from the method of randomized blocks, we choose a block of land large enough to accommodate the number of varieties we are comparing, and then lay out the experiment over a number of such blocks, the varieties being randomly allocated to the plots within a block, one to each. So long as there is not too much variation over a block, we may expect a fairly precise experiment by using the analysis of variance procedure to separate off the variation between variety means and between block means from the residual variation, which latter is used to measure the error of the experiment. In doing this we are tacitly assuming the block itself to be reasonably uniform in its fertility, although the mean levels of the blocks may differ greatly. If this is not so the variation in fertility will be reflected in our error. Now in some cases, particularly with relatively permanent experiments on, say, tea or rubber, we are quite likely to have data available consisting of the yields of these same plots in a year, or years, prior to the start of the experiment, at which time the treatment over the whole area has been uniform. The changes in fertility from one year to the next may be small enough for there to be a pronounced relationship between the yields of the previous year and those of the experimental year except in so far as the latter have become differentiated by the imposition of different treatments. This point can always be put to the test, for we can examine whether there is a significant association between the y (yield of the experimental year) amd x (that of the previous year). If significant, it is likely that we can, by the use of suitable statistical methods, use the x measures to improve the accuracy of the y measures. The methods will depend upon having an x corresponding to every y, and the

results will have their usual interpretation if it is clear that the x-measures have not themselves been affected by any treatment differences imposed on the y-measures, as will obviously be the case in the present illustration.

Previous uniformity trials are not, however, always available, and for many purposes they represent a luxury that cannot be afforded. Another method is sometimes possible. Let us suppose that the experiment is designed to test the effects of different manurial dressings, to be applied in the spring, on autumn-sown wheat. Suppose that at a time when the plants have weathered the worst of the winter and have become well-established, but before the spring dressings are applied, a count is made of the plant population on each plot. It may not be possible to do this fully, but a carefully-conducted sampling will at any rate yield a series of x-measures, one for each plot, x being number of plants. This figure may be a very good reflection of the fertility differences between the plots, and it is obviously not affected by treatment differences. At harvest-time we obtain the y-measures (e.g. yield), and again have a set of related x-measures to take into account. We may instance another case from the domain of animal experimentation, notorious for its high errors under the ordinary forms of management. Enough homogeneous material for, say, a feeding trial with farm animals to give a reasonably accurate experiment is hard to come by. Take, for instance, a pig-feeding trial. We may have a number of litters, equal to the desired amount of replication, and large enough to provide an animal from each litter for each imposed treatment. In this case each litter would form a block, in the usual terminology, but it would not usually be possible to guarantee that all the animals from the same litter were uniform; they are likely to differ somewhat in their initial weight at the time the experiment starts. After a time the animals are weighed again, the difference between initial and final weights being taken as the measure of growth y, and we are then in a position to see how the y averages compare as between different feeding treatments. More accurate measures of growth are possible if the animals are weighed at regular intervals throughout the experiment. The amount of growth, however, will also depend on the weight at the initial stage, since the animals may well be at somewhat different stages in their growth-curves. If we bring into account as x-measures the initial weights, these being measures obviously unaffected by the differential feeding treatments, we may find that we can improve the accuracy of the experiment by adjusting the rates of growth to correspond to equal initial weights.

We have considered in the foregoing paragraphs a number of examples where the purpose of a covariance analysis will be to effect an additional reduction, where possible, in the experimental error, i.e. after as much reduction as can be effected by the adoption of a suitable design has been achieved. A different type of problem is also amenable to the same form of analysis. Broadly this will come about when we wish to make a more detailed study of the differential effects of treatments than when we are content with a single measure like the yield of a plot. In all such cases we shall have a number of measures, instead of only one, for each plot, and we shall be concerned with the same statistical problems that arise with a comparison of homogeneous samples of observations in which there are more variates than one. The difference arises from the fact that the data of a field trial provide a more complicated sample pattern. Instead of having clearly independent samples to compare after the manner of the previous section, or in other ways not so far dealt with, we have to consider cases where the individual members of the samples undergoing different treatments have restrictions imposed upon them; for example, in the randomized blocks design one member of each sample is assigned to each block of land. In other cases the relationship will be even more complex. Furthermore, in experiments where two or more sets of treatments are incorporated there is the question of interactions between these to be considered. All this makes the necessary calculations a matter of special consideration in each case. The problems to be dealt with may be those of regression, with an affected variate and one or more affecting variates, or of correlation, or they may involve multivariate analysis in the sense that we may wish to have a single measure of the effect of treatments derived from a number of experimental measures. Broadly this set of problems may be distinguished from those of the preceding paragraph by the fact that some, or even all, of the variates, instead of only the affected variates, may be under differential treatment in the experiment. Both problems may, of course, be interwoven in one and the same experiment,

i.e. we may be concerned with the correlation of two measures present in differential treatment combinations, but may wish also to reduce our error by taking into consideration the measures of an affecting variate on which the treatments are without effect. Clearly each problem must be considered on its merits, and guidance as to the validity of the proposed analysis can only come from a complete understanding of the statistical methods which are available, paying attention to any underlying assumptions which have been made.

Analysis of Variance and Covariance

Let us first consider the case of two variates, the measures being denoted by x and y. The first step is to construct an *analysis of variance and covariance*. By this is meant that we work out by the standard methods, depending on the nature of the experimental layout (1940), not only an analysis of variance for x, but also one for y, and finally an analysis of covariance for the joint variation of x and y. For the first two of these there is nothing new; the total S.S. is broken up into its component parts, parts like Blocks, or Rows and Columns, being regarded as elimination of heterogeneity other than treatments, leaving us free to concentrate on the S.S. for Treatments in comparison with the S.S. for Error. In some cases the S.S. for treatments is also broken up into parts. Even in those cases where the x-measures are not differentiated by the treatments imposed, we must still work out the S.S. for "treatments" in order to have calculations in parallel for both x and y. A new feature of the analysis is that of the covariance. In this it is the total *sum of products* (S.P.) of deviations from the respective means of the x and y measures that is broken up into the same corresponding parts as the S.S. for x and for y. The calculations are similar to those for a S.S., a product of an x with a corresponding y taking the place of an x^2 or a y^2 at all stages; since, however, this is a new feature it would be well to illustrate the calculations with simple numbers. This is done in Tables 7 and 8, where x denotes the yield of a plot in a preliminary year under uniformity trial conditions, y the corresponding yield in the experimental year when there were three varieties, A, B and C, and there are four blocks of plots.

Table 7.—Data for x and y

Blocks		Varieties			Block Totals	
		A	*B*	*C*		
1	*x*	54	51	57	162	
	y	64	65	72		201
2	*x*	62	64	60	186	
	y	68	69	70		207
3	*x*	51	47	46	144	
	y	54	60	57		171
4	*x*	53	50	41	144	
	y	62	66	61		189
Total	*x*	220	212	204	636	
	y	248	260	260		768

Calculations. First find the *total* S.P. Multiply 54 by 64, 51 by 65, 57 by 72, and so on for all 12 products. Adding the results we get 40990. To correct for the means we must subtract (636 · 768) 12, or 40704. This leaves 286, with 11 D.F., for the total sum of products of deviations from the respective means (which are actually 53 and 64). Since the means are whole numbers in this example it would be easier to subtract them first, getting deviations of 1 and 0, − 2 and 1, 4 and 8, and so on, and then add the products of the deviations. But the first method is the standard one, and is expeditiously carried through when a calculating machine is available. We now require the *blocks* S.P. Multiply 162 by 201, 186 by 207, and so on for the four blocks. The addition of the products gives 122904, which we divide by 3, since each block is a total of 3 plots, giving 40968. We then subtract the above correction factor of 40704.

Thus the blocks S.P. is 264, with 3 D.F. To obtain the *varieties* S.P. we multiply 220 by 248, 212 by 260 and 204 by 260 and add the results, obtaining 162720. This we divide by 4 giving 40680, since there are 4 plots to each treatment. We then subtract the correction factor of 40704. The varieties S.P. is then — 24, with 2 D.F. Now in blocks and varieties we have accounted for 264 — 24 of the total S.P. 286. There is left for error 286 — 240, or 46, with 6 D.F. Table 8 shows these results, together with those for the S.S. for x and for y.

Table 8.—Analysis of Variance and Covariance

	D.F.	(x^2)	(xy)	(y^2)	b	r
Blocks	3	396	264	252	0·667	0·836
Varieties	2	32	—24	24	—0·750	—0·866
Error	6	86	46	48	0·535	0·716
Total	11	514	286	324	0·556	0·701

From this table we can first see that had we been given the y measures alone it would certainly not have been possible to demonstrate significant differences between varieties. The M.S. are 84, 12 and 8 in the order blocks, varieties, error, with 3, 2 and 6 D.F. respectively. We note that there has been a significant reduction of error by arranging the experiment in blocks, for the V.R. is 84/8 = 10·5, with 3 and 6 D.F. This is beyond the 1 *per cent.* point of the V.R. table. The V.R. for varieties is 1·5 with 2 and 6 D.F., the probability of this result being greater than 0·2. Let us now see what we can learn by taking the x measures into account. Let us first examine the general relationship between x and y. Had this been a homogeneous set of 12 pairs of observations we should have measured the relationship by means of the regression coefficient of y on x, calculated as $(xy)/(x^2)$, giving $b = 0{\cdot}556$, or by the correlation coefficient, which is (xy) divided by the geometric mean of (x^2) and (y^2), yielding $r = 0{\cdot}701$. These coefficients are significant (Fisher and Yates, Table VI), the value approaching the 1 *per cent.* level, for which r should be 0·708 with 10 D.F. This suggests that it is worth while going on. But the data are not homogenous. We may expect, for example, that if the fertilities of certain blocks are low, and of others high, the x and y measures may be positively related through this cause alone. Before considering how to adjust the y measures for their different values of x we have to determine the *net* relationship, freed from the possible effect that the blocks may have, and also from that of the varieties. In the latter case the fact that A, B and C represent the *same* varieties for the x measures, while the y means may differ if the yielding capacities of A, B and C are different, may lead to a lowering of the total correlation. To obtain the net relationship, then, we should determine the b (or the r) from the *error* line of the analysis of variance and covariance. In Table 8 all the b's (and r's) are given, although in practice we deal with the error line only and confine ourselves to b in this type of analysis. In testing significance we should remember that the D.F. are 2, 1 and 5 respectively for blocks, treatments and error residuals. We notice that b is positive for blocks and larger than the *total* b, and negative and large for varieties (this being brought about by the fact that by chance the x variety totals go down from A to C, whereas in the y-measures the figures for varieties B and C are higher than for A). Neither coefficient, however, is significant, in view of the small number of D.F. The *error* b is 0·535, only slightly lower, as it happens, than the *total* b, and it corresponds to an r of 0·716. With 5 D.F., however, we should require an r of 0·754 for significance at the 5 *per cent.* point. As a matter of fact the data were cut down to a few blocks (and adjusted slightly to give easy calculations), and so we need not be unduly disturbed at the lack of significance. An r of this magnitude would, in fact, have been significant had the D.F. been 6 instead of 5. We may, therefore, go on to demonstrate the next stage with the same data.

Analysis of Variance on Adjustment for x

What one does in practice is to ignore all these calculations that have been made to determine various b's and r's, and to concentrate at the first stage solely on the error line of the left-hand side of Table 8. We then test for the significance of the error regression as done in Table 5.

$(xy)^2/(x^2)$ is $46^2/86 = 24{\cdot}6$. This is (b^2), with 1 D.F., and (d^2) is then $48 - 24{\cdot}6 = 23{\cdot}4$, with 5 D.F. The M.S. is $23{\cdot}4 \cdot 5 = 4{\cdot}68$, and the V.R. is $24{\cdot}6/4{\cdot}68 = 5{\cdot}26$, with 1 and 5 D.F. This value lies between 4·06 for the 10 *per cent.* point and 6·61 for the 5 *per cent.* point (Fisher and Yates, Table V), and is thus not significant at the conventional standard (this is equivalent to the above test of $r = 0{\cdot}716$). In the ordinary way we might in consequence decide at this point that it was not worth while to go on making adjustments to the y's for the unequal values of x. But 5·26 is quite a substantial figure, and, as mentioned above, the D.F. are limited for the sake of providing an easy example. Our purpose here is to illustrate the further stages, and we need therefore make no apology for going on with the calculations. The caveat just entered is intended for the reader to apply in the course of his work, and so avoid unnecessary computation.

The variety means for x are 55, 53 and 51, with general mean 53, while for y they are 62, 65 and 65. The meaning of the regression coefficient being $+ 0{\cdot}535$ is that we should subtract from y 0·535 for every unit that x is above 53, and add 0·535 for every unit that x is below 53. The calculations in this case are very easy. For variety A we take $62 - 1{\cdot}07 = 60{\cdot}93$; for variety B the 65 is left unaltered (because x is 53), and for variety C we take $65 + 1{\cdot}07 = 66{\cdot}07$. The result has been to spread the y-measures out more than formerly, and we now infer that the effect of varietal differences is to be judged more accurately on these new estimates, which allow for the unequal values of x, than when the y-measures alone are available. The calculations are illustrated in Fig. 1. The values of the y- and x-means have here been plotted for the

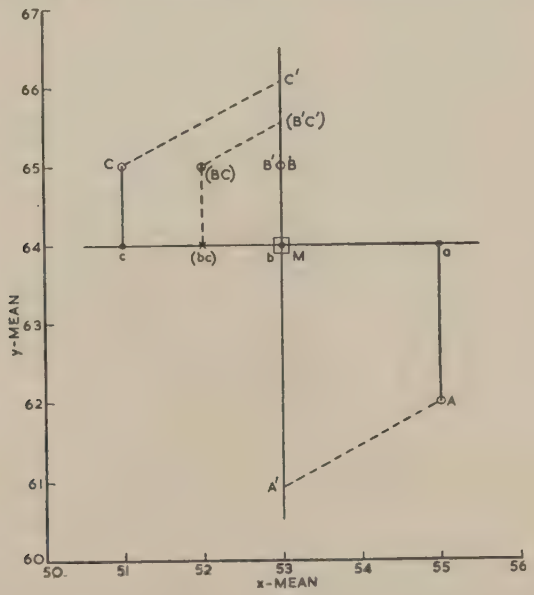

FIG. 1.—Adjustment of variety means of y for unequal values of x.

three varieties, giving the points A, B and C. Vertical and horizontal lines have been drawn through the general means 64 and 53, intersecting in M. If the feet of the perpendiculars from A, B and C on to the horizontal line through $\bar{y} = 64$ are called a, b $(= M)$ and c, then it is clear that the effect of varieties, unadjusted for x, is obtained from the deviations Aa, Bb and Cc from this horizontal line. In adjusting the y-means for the unequal values of x we move A and C along the dotted lines (having the slope $b = 0{\cdot}535$) till they meet the vertical line through

$\bar{x} = 53$ in the points A' and C' (B does not require adjustment, and so $B' = B$). What we now assert is that the effect of varieties is more accurately assessed by using the y-values corresponding to the points A', B' and C'. These are the values 60·93, 65 and 66·07 calculated above.

We now require a test of whether the adjusted varietal means are significantly different or not. A convenient method which requires the minimum of calculation, and which involves the ordinary variance ratio test, is one which is based on the test we have already described for the difference between the estimates of regression from two independent samples (see Table 6). We utilize the calculations of the left-hand side of Table 8, ignoring the line *blocks* (in a Latin square experiment we should equally ignore *both* of the lines for rows and columns). The results are shown in Table 9.

TABLE 9.—ANALYSIS OF RESIDUAL VARIANCE

	D.F.	(x^2)	(xy)	(y^2)	(b^2)	(d^2)	Diff.	D.F.
Varieties	2	32	−24	24			44·5	2
Error	6	86	46	48	24·6	23·4		5
Varieties + Error ..	8	118	22	72	4·1	67·9		7

The calculations are very simple. Just as for the error line (as shown at the beginning of this section), $(b^2) = 24·6$ and $(d^2) = 23·4$, so for the line varieties + error we have $(b^2) = 22^2/118 = 4·1$ and $(d^2) = 72 - 4·1 = 67·9$. The two values of (d^2) have 5 and 7 D.F. respectively, and their difference, 44·5, is written in the *varieties* line and accounts for the remaining 2 D.F.

We now have a S.S. of 44·5, with 2 D.F., for varieties adjusted for x, and a S.S. of 23·4, with 5 D.F., for residual variation. The mean squares are 22·25 and 4·68, and the variance ratio is therefore 4·75. We recall that, unadjusted, the variance ratio for varieties was only 1·5. The rise to 4·75 is a good indication that unequal values of x tended to obscure the varietal differences. Now 4·75, for D.F. 2 and 5, lies between 3·78 for the 10 *per cent.* significance level and 5·79 for the 5 *per cent.* level (Fisher and Yates, Table V). The effect cannot, therefore, be judged significant on the conventional standard, but again we may say that this is largely due to cutting down the data to four blocks for the purposes of easy illustration. With only 2 more D.F. for residual error a ratio of 4·75 would have been significant. The increase in precision that has been obtained by bringing the x-variate to account is seen by comparing the earlier error M.S. of 8, with 6 D.F., with the new figure of 4·68, with 5 D.F.; the accuracy has been nearly doubled.

In case any reader calculates from the points A', B' and C' in Fig. 1 that the sum of squares of their deviations from M (when multiplied by 4, the number of blocks) is 58·8, it should be pointed out that this is not the figure which should appear as a "difference" in Table 9, but that, nevertheless, the "difference" 44·5 is correctly used in the manner indicated above to measure the effect of varietal differences adjusted for x. Note also that the difference between Tables 6 and 9 is that in the former we had two clearly independent samples and were therefore able to add the values of (d^2) to get a pooled residual; also what we were after was to compare the two values of b. In Table 9 we have only one value of (d^2), namely the one from the error line, that we can safely use as a measure of residual variation, and so we compare the whole of the remainder with it. Included in the "difference", of course, is a component which would, if isolated, measure the difference between the b's calculated from the varieties and error lines (− 0·750 and 0·535), but it is seldom necessary to examine this component by itself. The remainder is a measure of the variation of the variety means round the regression line ($b = -0·750$) fitted to them.

Finally, we require the standard errors of the adjusted y-means for varieties A, B and C, and the standard error of the difference between any two of these. This is for final tabulation purposes, and also to investigate more closely what has contributed to a nearly significant variance ratio for varieties as a whole, when corrected for unequal values of x. For the calculation we shall use the residual M.S. of 4·68 to determine the variance of a mean of 4 determinations, namely $4·68/4 = 1·17$, and also the variance of b, which is $4·68/86 = 0·0544$ (see Tables

4 and 8). Now the adjusted y-mean for variety A is $62 - 2b$, while that for variety C is $65 - 2b$, both consisting of two independent parts. Thus the variances for these two means are equal, and are calculated as 1·17 plus *four times* 0·0544 (the multiplier is 2^2), giving 1·30, of which the square root gives the standard error of the adjusted y-means as 1·18. For variety B the y-mean is unadjusted, since x is 53, and its standard error is therefore $\sqrt{1·17} = 1·08$. Let us now consider the difference between two adjusted y-means, say 60·93 for A and 66·07 for C, noting that these are not independent, because b has entered into their calculation. The *difference*, 5·14, is in fact made up of two independent parts, namely (*a*) 3, the difference between the uncorrected y-values 65 and 62, and (*b*) 2·14, being four times the value 0·535 of b, the figure 4 arising from the difference between the corresponding x-values 55 and 51. The first part, being the difference of two independent means of 4 values each, has variance $2 \times 1·17 = 2·34$, while the variance of the second part is $16 \cdot 0·0544 = 0·87$, 16 being the square of the multiplier 4. The variance of the difference 5·14 is then $2·34 + 0·87 = 3·21$, and the standard error $\sqrt{3·21} = 1·79$. This, the largest of the three differences, is 2·9 times its standard error, whereas the t-test (Fisher and Yates, Table III) shows that a ratio of under 2·6 is required for significance at the 5 *per cent.* point with 5 D.F., for a difference selected at random. There appear, therefore, to be grounds for asserting that variety C is better than variety A.

Unfortunately, this calculation has to be made afresh for each pair of means compared, for the second part of the difference depends on the difference of the corresponding x's. In addition, we have to be a little careful with our deductions. Strictly speaking, we should not claim a difference as significant when the variance ratio of Table 9 failed to show up as significant. We chose the difference between A and C as the largest of the three possible differences, and then applied a criterion to it which is strictly only applicable to a difference chosen at random. This situation must be familiar to readers who have had experience with ordinary experiments. The present data can be used to indicate a way round the difficulty. We observe that the uncorrected means for A, B and C are 62, 65 and 65. In the analysis of variance of the y-values we would be entitled to split up the S.S. for varieties into a part for the difference between A and the mean of B and C, i.e. between 62 and 65, and a remaining part for the difference between B and C (in this case zero). These parts will have 1 D.F. each. The general calculation will be illustrated from the totals (of 4 plots) 248, 260, 260 (see Table 7). The calculation for the first part is $248^2/4 + 520^2/8 - 768^2/12 = 24$; for the second part $260^2/4 + 260^2/4 - 520^2/8 = 0$ (more simply $(260 - 260)^2/8$). Now the specific question of the possible significance of the difference between A and the mean of B and C can be answered if we make a similar division in both the (x^2) and (xy) parts of the varieties line in Table 8, and then proceed with the calculations of Table 9 for this part of "varieties" in association with error. This is done in Table 10:

TABLE 10

	D.F.	(x^2)	(xy)	(y^2)	(b^2)	(d^2)	Diff.	D.F.
A (B C)	1	24	−24	24			44·2	1
Error	6	86	46	48	24·6	23·4		5
A (B C)+Error	7	110	22	72	4·4	67·6		

The variance ratio is $44·2/4·68 = 9·44^{**}$, with 1 and 5 D.F. The 5 *per cent.* point is 6·61 and the 1 *per cent.* point 16·26. We have thus demonstrated that the average performance of B and C is superior to that of A when allowance is made for the unequal values of x. Without this allowance the variance ratio would be $24/8 = 3$, as against the former 1·5 for the whole of varieties. With 1 and 6 D.F. this value of 3 is not significant.

There is a special feature peculiar to this calculation, namely that the values of B and C are equal. But the explanation given should serve to indicate how the calculation is done in general. Wherever, as in Table 10, we get a set (x^2), (xy) and (y^2) for only 1 D.F. the numbers must be such that the r calculated from them is either $+1$ or -1. For we have only two pairs of values; here the means are 55 and 52 for x and 62 and 65 for y. With x going down from A to (B C) and y going up we must have a correlation of -1. The only possibilities yielding a different

result would be x and y going up together or down together, in which case the correlation would be + 1. In the present case (x^2), (xy) and (y^2) are all equal in magnitude, but in general b will not be unity, the only condition holding being $(xy)^2 = (x^2)(y^2)$.

Because the above variance ratio 9·44 has 1 and 5 D.F. its square root 3·07 must be the value of t for the ratio of the difference between A and the mean of B and C, both corrected for unequal x, to its standard error. These values are 60·93 and 65·54 and the difference 4·61. This shows that the standard error of the difference 4·61 must be 4·61/3·07 or 1·50. Without doing the calculations of Table 10 we could have worked out this standard error directly. Since the corrected value for A is $62 - 2b$ and for the mean of B and C $65 + b$ (since x for the mean of B and C is 52), the difference between the two means is $(65 - 62) + 3b$. 65 is the mean of 8 values and therefore has the variance $4{\cdot}68/8 = 0{\cdot}58$, while 62, as we saw above, has variance 1·17. The variance of the third term is $9 \times 4{\cdot}68/86 = 0{\cdot}49$. Adding these values we have 2·24 for the variance of the difference between A and the mean of B and C, giving a standard error of $\sqrt{2{\cdot}24} = 1{\cdot}50$. The means compared are shown in Fig. 1 as the points A' and $(B'\ C')$. Either method, i.e. the modification of the analysis of residual variance table or the calculation of the appropriate t, may be adopted with a single degree of freedom effect.

A similar modification of the analysis table should always be resorted to when two or more interacting sets of treatments are included in an experiment. Just as in the ordinary analysis of variance the treatments S.S. is divided into component parts to test separately for the significance of the direct effects and also of the various interactions, so we shall be interested in testing separately for these effects when allowance has been made for the unequal values of x, assuming that an analysis of the form of Table 8 has shown a significant error regression. In such a case Table 8 would be worked out in full, i.e. the quantities (x^2), (xy) and (y^2) would be determined for each component of the treatments. Then, instead of including the whole of the treatments in the calculations of Table 9 we should instead set out in the form of Table 10 the chosen component of the treatment effect against the error. The test, by means of a variance ratio, of the "difference" against the residual M.S., would then tell us whether the particular component was significant after correction for x. If the component had more than 1 D.F. we should then be interested, if the effect was significant, in calculating the standard errors of the differences between the various adjusted y-means, by the method previously described.

Example

The analysis of the previous sections, when pruned of the somewhat lengthy explanatory accompaniment, and of certain alternative presentations of the tests involved, boils down to a straightforward and relatively simple series of operations. These will now be illustrated on the data of a full-scale experiment. In an experimental piggery, arranged for individual feeding, there were five pens, with room for six pigs in each pen. Six young bacon pigs, three being hogs and three gilts, were selected from a single litter and allotted to one of the pens. The remaining pens were made up in the same way from other litters. The 30 pigs were all weighed individually at the start of the experiment. Three different feeding treatments were then introduced, each being given to a hog and a gilt from each pen. Denoting these treatments by A, B and C, the difference was that A contained crude protein ranging from 17·5 *per cent.* to 12·2 *per cent.* as the experiment proceeded, while the ranges for B and C were 22·1 *per cent.* — 16·9 *per cent.* and 26·8 *per cent.* — 21·7 *per cent.* respectively. The pigs were individually weighed each week for 16 weeks. The purpose of the experiment was to compare the pigs resulting from the three treatments, and to examine also the possible sex differences. For the purposes of the present example we shall only consider the measured growth-rates. For each there were seventeen equally-spaced weights, and on the assumption that the growth was linear a simplified version (because of the equal spacing) of the method described in the second section of this bulletin enabled a regression coefficient to be calculated as a measure of growth in pounds per week. It is these 30 figures, denoted by g, which fall to be analysed. To improve their precision we have the initial weights, w_0, before differential feeding started, to take into account. This is particularly necessary in the present case, because it was thought that more representative samples from each litter could be obtained if three light and three heavy pigs were chosen. Normally these were

chosen to be either all hogs or all gilts, but this was not always possible, and we note also that there was an odd number of pens. It was therefore suspected that the choice would influence the conclusions to be drawn between the sexes, and also possibly between treatments. Some allowance, therefore, had to be made for initial weight.

The initial weight w_o in pounds, and the linear growth-rate g in pounds per week, are given in Table II.

TABLE II

Pen		A		B		C		Total
		H	G	H	G	H	G	
I	w_o	38	48	39	48	48	48	269
	g	9·52	9·94	8·51	10·00	9·11	9·75	56·83
II	w_o	35	32	38	32	37	28	202
	g	8·21	9·48	9·95	9·24	8·50	8·66	54·04
III	w_o	41	35	46	41	42	33	238
	g	9·32	9·32	8·43	9·34	8·90	7·63	52·94
IV	w_o	48	46	40	46	42	50	272
	g	10·56	10·90	8·86	9·68	9·51	10·37	59·88
V	w_o	43	32	40	37	40	30	222
	g	10·42	8·82	9·20	9·67	8·76	8·57	55·44
Total	w_o	205	193	203	204	209	189	1203
	g	48·03	48·46	44·95	47·93	44·78	44·98	279·13

The form of analysis is that appropriate to a randomized block experiment (the pens being the blocks) containing six treatment combinations of the three foods and the two sexes. The total variation (or co-variation) is therefore divided up into a component for pens with 4 D.F., a component for "treatments" with 5 D.F., further subdivided into food (2 D.F.), sex (1 D.F.) and interaction of food and sex (2 D.F.), and finally a component for error with 4 × 5 or 20 D.F. The analysis is done separately for the S.S. of initial weight (w_o^2), for the S.S. of growth-rate (g^2), and for the S.P. of initial weight and growth-rate ($w_o g$). The resulting figures are shown in Table 12, in which the ordinary analysis of variance and covariance has been bordered by rows for the various combinations of treatments and error, and by columns to show (b^2) and (d^2) and the various "differences", on the lines of Table 9.

TABLE 12

	D.F.	(w_o^2)	$(w_o g)$	(g^2)	(b^2)	(d^2)	Diff.	D.F.
Pens	4	605·87	39·905	4·8518				
Food	2	5·40	−0·147	2·2686			2·3365	2
Sex	1	32·03	−3·730	0·4344			1·2594	1
Interaction ..	2	22·47	3·112	0·4761			0·0978	2
Error	20	442·93	39·367	8·3144	3·4989	4·8155	(0·2534)	19
Total	29	1108·70	78·507	16·3453				
Food + Error ..	22	448·33	39·220	10·5830	3·4310	7·1520		
Sex + Error ..	21	474·96	35·637	8·7488	2·6739	6·0749		
Inter. + Error	22	465·40	42·479	8·7905	3·8772	4·9133		

$b = 39·367/442·93 = +0·089$ lb. per week.

For each line in which it is calculated (b^2) is $(w_o g)^2/(w_o^2)$, while (d^2) is the result of subtracting this quantity from (g^2), and the calculation is done for error and for the lines giving food, sex

and interaction in combination with error. When (d^2) for error is subtracted from one of the others the resulting "difference" is placed in the appropriate line in the top part of the table. The M.S. for error, got by dividing (d^2) by 19, is placed in brackets, and the D.F. for the various variance-ratio comparisons that can be made are placed on the extreme right.

All the necessary calculations are then displayed in one table, but in practice we first work out (b^2), (d^2) and M.S. for the error line, in order to test whether the regression is significant before proceeding with the rest of the calculation. The variance ratio is $3{\cdot}4989/0{\cdot}2534 = 13{\cdot}81^{**}$, with 1 and 19 D.F., and is significant at the 1 *per cent.* point (the 0·1 *per cent.* point is 15·08). By taking account of initial weights the M.S. for error has been reduced from $8{\cdot}3144/20 = 0{\cdot}4157$ to 0·2534. The variance ratios for the various treatment effects may now be calculated. That for food is $1{\cdot}1682/0{\cdot}2534 = 4{\cdot}61^{*}$ with 2 and 19 D.F., significant at the 5 *per cent.* point (the 1 *per cent.* point is 5·93). For sex we have $1{\cdot}2594/0{\cdot}2534 = 4{\cdot}97^{*}$, significant at the 5 *per cent.* point, while for interaction the figure is $0{\cdot}0489/0{\cdot}2534 = 0{\cdot}19$, insignificant.

We have thus demonstrated significant differences between the growth-rates for the three foods, on the average of the two sexes, and for the two sexes, on the average of the three foods, and also find that because of the absence of a significant interaction, these two effects tell us all that can be learnt from the data so far analysed. By calculating the various M.S. from the (g^2) column of Table 12, the reader may verify that the effects of food and sex (and of interaction) are insignificant if no account is taken of the initial weights.

The growth-rate means may be displayed in the following short tables, in which we use $b = +8/90$ approximately:—

TABLE 13.—RESULTS

Food	*A*	*B*	*C*	Mean
mean w_o	39·8	40·7	39·8	40·1
unadjusted mean *g*	9·649	9·288	8·976	9·304
adjusted mean *g*	9·676	9·235	9·003	9·304
S.E.	0·159	0·160	0·159	

Sex	*H*	*G*	Mean
mean w_o	41·13	39·07	40·1
unadjusted mean *g*	9·184	9·425	9·304
adjusted mean *g*	9·092	9·517	9·304
S.E.	0·132	0·132	

The first two lines call for no comment, being the means of 10 and 15 observations respectively, obtained from the column totals of Table 11. To illustrate the calculation of an adjusted mean let us take food *A*; w_o is 0·3 *below* the mean 40·1; we therefore *add* $0{\cdot}3 \times 8/90$ or 0·027 to 9·649, getting 9·676. The standard errors (S.E.) of the adjusted means are calculated as previously described. The first, for example, is $\sqrt{\{0{\cdot}2534(0{\cdot}1 + 0{\cdot}09/442{\cdot}93)\}}$, while that for the adjusted *H* mean is $\sqrt{\{0{\cdot}2534(0{\cdot}0667 + 1{\cdot}0609/442{\cdot}93)\}}$. In the present case these figures differ very little from the figures 0·159 and 0·130 applicable to unadjusted means of 10 and 15 observations respectively.

We know from the analysis that there are significant differences between the mean *g*'s for *A*, *B* and *C*, and between *H* and *G*. We need go no further with the *H*, *G* comparison, but if we wish to test specifically for the difference between *A* and *B*, say, we must note that the estimated variance of this difference is obtained by multiplying the error M.S. 0·2534 by

$$\frac{2}{10} + \frac{0{\cdot}9^2}{442{\cdot}93}.$$

The first part arises from comparing 2 means of 10; in the second 0·9 is the difference between the mean w_o's for *A* and *B*, and 442·93 is the error S.S. for w_o. The calculation gives us 0·2018 (the effect of the second term being very small in this case), and so the estimated variance is $0{\cdot}2534 \times 0{\cdot}2018 = 0{\cdot}0511$, of which the square root is 0·226. The difference between the

mean g's is 0·441 and t is 0·441/0·226 = 1·95, with 19 D.F. The 5 *per cent.* point is 2·09 and we therefore judge this difference to fall short of significance. We conclude that the drop in mean g from A to C is significant (because of the results of Table 12) but that the differences between A and B, and, of course, between B and C, are not individually so.

Note on calculation.—The only part of the calculation which has not been explained above is that leading to the figures in the analysis of variance and covariance in Table 12. To obtain the S.S. we follow the method given in the earlier bulletin (1940, pp. 22 23). For the S.P. determine first the correction factor (1203) (279·13), 30 11,193·113. Then sum all products of w_o and g (the first is 38 × 9·52). This gives 11,271·62, and on subtracting the C.F. we get 78·507 for the *total* line of Table 12. For "pens" sum the products of the treatment totals (the first is 269 × 56·83), divide by 6 and subtract the C.F., obtaining 39·905. For "treatments" sum the products of the pen totals (the first is 205 × 48·03), divide by 5 and subtract the C.F., giving − 0·765 for the total of food, sex and interaction. "Error" is now obtained by difference as 78·507 39·905 − (− 0·765) = 39·367. Now to break up "treatments" note that the A, B and C totals (adding H and G) are, from the bottom row of Table 11, 398, 407 and 398 for w_o, and 96·49, 92·88 and 89·76 for g. Sum the products of these numbers, divide by 10 and subtract the C.F., obtaining - 0·147 for "food". Note again from Table 11 that the H and G totals (adding A, B and C) are 617 and 586 for w_o and 137·76 and 141·37 for g. Since only two totals are involved in this case for each variate, a quick method is to multiply (617 − 586) by (137·76 − 141·37) and divide by 30, the total number of observations. The answer is − 3·730, the appropriate S.P. for "sex". "Interaction" is obtained by difference, i.e. we calculate

$$-0{\cdot}765 - (-0{\cdot}147) - (-3{\cdot}730) = -0{\cdot}765 + 0{\cdot}147 + 3{\cdot}730 = 3{\cdot}112.$$

Or it may be got by calculating the pairs of $H - G$ differences for the three foods (the first, for example, are 12 and − 0·43). Summing these three products, dividing by 10 and subtracting the "sex" S.P. (− 3·730) we obtain 3·112.

Difference Between Regressions in Two Correlated Samples

We saw earlier that the method of adjusting the observations in, say, a field trial for values of an affecting variate depended essentially upon having a test for the difference between the regression coefficients that can be calculated from two independent samples. The object of this section is to consider the modifications that must be made when the samples are no longer independent. We shall consider here the case of two homogeneous samples. The data used in Table 1, and again in Table 6, consisted of the yield at harvest of a number of bean varieties together with the number of shoots in March. If we add in an extra pair of values we have five values of the y series and five of the x series. Let us now suppose that there were five *pairs* of plots in the experiment, that each pair were sown with the same sort of bean, but that one plot out of each pair was given a different manurial treatment from the other. We should then have two series of y's (y_1 and y_2) and two series of x's (x_1 and x_2), and the problem is to compare the relation between yield and shoot number as between the series denoted by the subscripts 1 and 2. The assembled data chosen for analysis are given in Table 14, in which a pronounced correlation can be seen to exist between y_1 and y_2 (r = 0·96), while the r between x_1 and x_2 is 0·69.

Table 14

x_1	y_1	x_2	y_2
173	10·7	152	10·2
148	9·8	137	9·4
153	9·5	127	8·6
122	8·4	119	7·9
129	9·1	140	8·9
Total 725	47·5	675	45·0
Mean 145	9·5	135	9·0

From this table we may readily work out all the S.S. and S.P., obtaining the figures shown in Table 15:—

TABLE 15

	(*xx*)	(*xy*)		(*yy*)
(1, 1)	1642	66·2	7·7	2·90
(1, 2)	−706	−58·5		−2·81
(2, 1)	−706	−36·6	4·9	−2·81
(2, 2)	638	41·5		2·98
Total	868	12·6		0·26

$\bar{b} = 12{\cdot}6/868 = 0{\cdot}0145$

$12{\cdot}6^2/868 = 0{\cdot}1829$ (1 D.F.) *mean regression*

$1642 \times 638 - 706^2 = 549{,}160$

$549{,}160\, b_1 = 638 \times 7{\cdot}7 + 706 \times 4{\cdot}9 = 8372 \qquad b_1 = 0{\cdot}01525$

$549{,}160\, b_2 = 706 \times 7{\cdot}7 + 1642 \times 4{\cdot}9 = 13482 \qquad b_2 = 0{\cdot}02455$

Difference 5110

$\dfrac{5110^2}{868 \times 549160} = 0{\cdot}0548$ (1 D.F.) *difference between regressions*

$0{\cdot}26 - 0{\cdot}1829 - 0{\cdot}0548 = 0{\cdot}0223$ (2 D.F.) *residual from separate regressions.*

The column headings denote that the first relates to S.S. or S.P. for the two x series; the second to the S.P. between the x and y series, and the third to the S.S. or S.P. for the two y series. In the first line we have (x_1^2), (x_1y_1) and (y_1^2); in the second (x_1x_2), (x_1y_2) and (y_1y_2); in the third (x_2x_1), (x_2y_1) and (y_2y_1); and in the last (x_2^2), (x_2y_2) and (y_2^2). *Note that in the second and third line the S.P. are written with reversed signs*, i.e. the items in the first column are (x_1^2), $-(x_1x_2)$, $-(x_2x_1)$ and (x_2^2). The columns have been summed, also the bracketed pairs from the (xy) column.

The remainder of the calculation is shown below the table. The regression coefficient $\bar{b}$ of $y_1 - y_2$ on $x_1 - x_2$ is calculated in the usual way from the total line, and the S.S. for $y_1 - y_2$, namely 0·26, with 4 D.F., will contain a component $(\bar{b}^2)$ due to this regression, amounting to 0·1829, with 1 D.F. The remainder, 0·0771, is the S.S. of residuals of $y_1 - y_2$ from the fitted line of regression. The next three lines show how b_1 and b_2 can be separately calculated. Next, the S.S. of residuals will contain a component depending on the difference between b_1 and b_2. This component, calculated as shown, amounts to 0·0548, with 1 D.F. Subtracting this from 0·0771 we have our final figure 0·0223, with 2 D.F., which is the S.S. of the difference of residuals of the y's from the two lines of regression, with coefficients b_1 and b_2. There will be a factor 2 in all these S.S., since they are calculated from the differences of the paired values.

The break-up is shown in analysis of variance form in Table 16:—

TABLE 16

	D.F.	2 S.S.	M.S.	V.R.
Mean regression	1	0·1829		
Difference between regressions	1	0·0548	0·0274	4·9
Residual	2	0·0223	0·0056	
Total	4	0·2600		

The variance ratio for the difference between the regressions is 4·9, with 1 and 2 D.F. This is not significant, the 5 *per cent.* point being 18·5, and we conclude that the two series do not differ significantly in the way in which yield depends upon shoot number in March, after allowing for the correlation between y_1 and y_2. Had we ignored the correlation and used the method exemplified in Table 6 we should have found $b_1 = 0{\cdot}040$, $b_2 = 0{\cdot}065$ and V.R. = 3·3, with 1 and 6 D.F., again insignificant and lower than the V.R. we have just found. But the residual mean

square would have been 0·0852, with 6 D.F., compared with our 0·0056 with 2 D.F. The remaining D.F. are accounted for by the differences between the five sums $y_1 + y_2$ of the paired values. If we add the last column of Table 15 with the centre values given their correct signs (positive) we get 11·5. Half of this, namely 5·75, is the S.S. for the differences between the five pair means, with 4 D.F. We note that 5·75 plus 0·13 (see Table 16) is 5·88, which is $(y_1^2) + (y_2^2)$ with 8 D.F., as used when we follow the method of Table 6. The method we now use is similar to the "Student" test of the difference between means in paired comparisons, and, as in that test, we have here eliminated the differences between the values of $y_1 + y_2$, and thereby the effect of the correlation between the samples.

We may test b_1 and b_2 separately for significance by calculating their standard errors. The base of this operation is the residual S.S. $s^2 = 0{\cdot}0056$. To obtain the estimated variance of b_1 we multiply s^2 by (see Table 15)

$$\frac{2 \times 638}{549{,}160} = 0{\cdot}002324$$

and the standard error of b_1 is then 0·0036. The ratio of b_1 to its standard error is 4·2, just under the value of 4·3 required for significance at the 5 *per cent.* point. For b_2 we replace 638 above (which is (x_2^2)) by 1642 (which is (x_1^2)) to get a new multiplier of s^2, namely 0·005980. We then find the S.E. of b_2 to be 0·0058, so that $t = 4{\cdot}25$. The coefficients are therefore at least strongly suggestive of an association between yield and stem number in both samples which is not due to the correlation existing between the y_1 and y_2 series. We could equally well determine the estimated covariance of b_1 and b_2 by replacing the above 638 by 706 (which is (x_1x_2)) to get a third multiplier of s^2, namely 0·002571. Then the variance of $b_1 - b_2$ is the variance of b_1 plus that of b_2, less twice the covariance of b_1 and b_2, i.e. it is s^2 multiplied by 0·002324 + 0·005980 − 2(0·002571) = 0·003161. We then get 0·0042 as the standard error of $b_1 - b_2$. But this test has already been carried out in Table 16, for the variance ratio 4·9 is just the square of the ratio of 0·0093 to 0·0042.

For the further development of the test of this section as applied to more samples than two, and for more than one affecting variate, the reader is referred to a paper by A. H. Carter (1949, *Biometrika* **36**, 26). The section immediately following helps to explain the calculations which were based on Table 15.

Regression with Two Affecting Variates

The general principles of the analysis of covariance procedure have now been given, but the illustrations only show how adjustment is carried out for a single affecting variate. Obviously there is nothing in the method which cannot be applied to cases where there are two or more affecting variates, but we must first see how to measure regression (or correlation) in the general case. Let us begin with the case of two such variates, which we shall call the x_1 and x_2 series, retaining y for the dependent or affected variate. Following on the lines of the second section of this bulletin, let us first consider the homogeneous case, illustrating by means of an arithmetical example with simple numbers. Take the data of the first two columns of Table 14, reading y for y_1, and add figures for the number of stems at harvest (now to be called x_2) from the same experiment. We then get the numbers shown in Table 17 where, without more ado, we have calculated the three S.S. and the three possible S.P. by the methods already described.

Table 17

x_1	x_2	y
173	123	10·7
148	99	9·8
153	111	9·5
122	88	8·4
129	94	9·1
Total 725	515	47·5
$\bar{x}_1 = 145$	$\bar{x}_2 = 103$	$\bar{y} = 9{\cdot}5$

$(x_1^2) = 1642$
$(x_2^2) = 786$
$(y^2) = 2{\cdot}90$
$(x_1x_2) = 1101$
$(x_1y) = 66{\cdot}2$
$(x_2y) = 42{\cdot}9$

We seek a relationship in which y is expressed linearly in terms of x_1 and x_2 jointly. Because there are two affecting variates there will be two *partial regression coefficients* to be estimated. Let these estimates be denoted by b_1 and b_2. They are derived as follows:—

$$b_1 = \frac{(x_1y)\,(x_2^2) - (x_2y)\,(x_1x_2)}{(x_1^2)\,(x_2^2) - (x_1x_2)^2} = \frac{66{\cdot}2 \times 786 - 42{\cdot}9 \times 1101}{1642 \times 786 - 1101^2} = 0{\cdot}06122$$

$$b_2 = \frac{(x_2y)\,(x_1^2) - (x_1y)\,(x_1x_2)}{(x_1^2)\,(x_2^2) - (x_1x_2)^2} = \frac{42{\cdot}9 \times 1642 - 66{\cdot}2 \times 1101}{78411} = -0{\cdot}03117$$

The *regression equation* is

$$Y = 9{\cdot}5 + 0{\cdot}061\,(x_1 - 145) - 0{\cdot}031\,(x_2 - 103)$$

from which, by inserting any pair of values of x_1 and x_2 from Table 17, we may obtain the calculated value Y which corresponds to the y associated with those values. Take the first line. x_1 is 173 and x_2 is 123. Then:—

$$Y = 9{\cdot}5 + 0{\cdot}061 \times 28 - 0{\cdot}031 \times 20 = 10{\cdot}59.$$

The remaining values are set out in Table 18 against the values of y.

TABLE 18

y	Y	$y - Y$	$(y - Y)^2$
10·7	10·59	0·11	0·0121
9·8	9·81	−0·01	0·0001
9·5	9·74	−0·24	0·0576
8·4	8·56	−0·16	0·0256
9·1	8·80	0·30	0·0900
Total 47·5	47·50	0·00	0·1854

We find that we have a residual or deviation S.S., which as before we may call (d^2), amounting to 0·1854. More exactly, to 4 decimal places, this should be 0·1846, as can be seen by utilising the four significant figures given for the b's. In obtaining this value we have used $\bar{y}$, b_1 and b_2, so that it has $5 - 3 = 2$ D.F. The estimate of the residual variance, which we call s^2, is then 0·1854/2, or 0·0927 (more exactly 0·0923). The square root, namely $s = 0{\cdot}304$, may be called the standard error appropriate to the regression equation.

We have now to consider what the appropriate test of significance is. We want to know whether there is a significant association between y and x_1 and x_2 jointly, i.e. whether the b_1 and b_2 are such that it is unlikely that their true values are both zero. Such a test is very similar to that obtained earlier in the case of a single affecting variate, and is most simply demonstrated by an analysis of variance, in which the S.S. (v^2), which is already known to be 2·9, is divided into a part (b^2) "due to" the regression, but having this time 2 D.F. (because of b_1 and b_2), and a remaining part (d^2) representing deviations from regression. The second part is already known to be 0·1846, and the first part must therefore be 2·9 − 0·1846 = 2·7154. The decomposition, and the remaining stages in making the test, are shown in Table 19:—

TABLE 19

	D.F.	S.S.	M.S.	V.R.
Regression ..	2	2·7154	1·3577	14·7
Residual ..	2	0·1846	0·0923	
Total	4	2·9000		

With 2 and 2 D.F. the 10 *per cent.* point for the variance ratio is 9·0 and the 5 *per cent.* point 19·0. On the conventional standard the regression must in this case be judged non-significant. This result does not really contradict that of Table 5, as we shall shortly see; the use of only a few

sets of observations for ease of illustration inevitably has the result that a high value of the V.R. is required for significance because of the small number of D.F. on which the residual error is based.

The question may now be asked whether it is necessary to calculate the Y values in order to obtain (d^2). The answer is that it is not. As in the case of simple regression we may calculate (b^2) directly from the S.S. and S.P. of Table 17. The numerator of (b^2) is

$$(x_2^2)\,(x_1y)^2 + (x_1^2)\,(x_2y)^2 - 2\,(x_1x_2)\,(x_1y)\,(x_2y) = 212{,}915{\cdot}1,$$

while its denominator is that of b_1 and b_2, namely $(x_1^2)\,(x_2^2) - (x_1x_2)^2 = 78{,}411$. We find, therefore, that (b^2) is 2·71754, whence (d^2) is 0·1846 by subtraction from $(y^2) = 2{\cdot}9$. Alternative expressions for (b^2) are $b_1\,(x_1y) + b_2\,(x_2y)$ and $b_1^2\,(x_1^2) + 2b_1b_2\,(x_1x_2) + b_2^2\,(x_2^2)$, but note that if these are used b_1 and b_2 must be used to more places than are quoted below Table 17 if (b^2) is to be correct to 4 decimal places.

It is usually of interest to test separately for the significance of the partial regression coefficients b_1 and b_2. One way to do this is to calculate their standard errors, as we did in the last section for the b_1 and b_2 defined therein. The base of the operation is s^2, whose value is 0·0923. The estimated variance of b_1 is obtained by multiplying s^2 by

$$\frac{(x_2^2)}{(x_1^2)\,(x_2^2) - (x_1x_2)^2} = 0{\cdot}01002$$

giving 0·000925, from which, by extracting the square root we have the standard error of b_1 as 0·0304. The ratio of b_1 to its S.E. is $t = 0{\cdot}061/0{\cdot}0304 = 2{\cdot}0$, with D.F. 2, which is not significant (the 5 *per cent.* point is 4·3). Similarly, to obtain the estimated variance of b_2 we multiply s^2 by

$$\frac{(x_1^2)}{(x_1^2)\,(x_2^2) - (x_1x_2)^2} = 0{\cdot}02094$$

giving 0·001933, of which the square root is 0·0440. t is then $0{\cdot}031 \div 0{\cdot}044 = 0{\cdot}7$, and is clearly non-significant.

A further test of value is to compare b_1 with b_2. The estimated covariance of b_1 and b_2 is obtained by multiplying s^2 by

$$-\frac{(x_1x_2)}{(x_1^2)\,(x_2^2) - (x_1x_2)^2} = -0{\cdot}01404.$$

Now the estimated variance of $b_1 - b_2$ is the variance of b_1 plus that of b_2, less twice the covariance of b_1 and b_2. It is thus obtained as s^2 multiplied by $0{\cdot}01002 + 0{\cdot}02094 + 2(0{\cdot}01404) = 0{\cdot}05904$. The estimated variance is therefore $0{\cdot}0923 \times 0{\cdot}05904 = 0{\cdot}005451$, of which the square root is 0·0738. The difference $b_1 - b_2$ being $0{\cdot}061 - (-0{\cdot}031) = 0{\cdot}092$, we obtain for t the value $0{\cdot}092/0{\cdot}0738 = 1{\cdot}3$, with 2 D.F. This is not significant.

It should now be recognized that the calculations given in the last section, both in connexion with Table 15 and in the working out of the standard errors, spring from the fact that the problem of that section is mathematically equivalent to working out the regression of the variate $y_1 - y_2$ on the variates x_1 and x_2. The reader is recommended to turn back and verify this for himself, with the aid of the formulæ of the present section.

A point of great practical importance is to know where to stop in regard to the number of variates which may possibly be related to the variate y. The calculations become tedious with more than two affecting variates, and it is important to have some means of proceeding in stages from one upwards, at each stage determining the significance of the coefficient due to the added variate. If insignificant it may be discarded and another chosen, unless it is found that the significance of the regression *as a whole* has been improved. For it is possible to have the individual partial regressions insignificant and yet for regression (from such a table as Table 19) to be significant. In the present example we saw earlier on that the regression of yield of beans at harvest (y) on number of bean shoots in March (x_1) was significant. Adding in the fifth pair of values from Table 17, namely $x_1 = 129$, $y = 0{\cdot}1$, the S.S. of Table 5 become 2·6690, 0·2310 and 2·9, with 1, 3 and 4 D.F. The mean squares are 2·6690 and 0·077, and the variance ratio 34·7**, with 1 and 3 D.F. This is now significant at the 1 *per cent.* point. Let us now determine

whether anything is gained by bringing in x_2. The calculations are very easy. From the figures just given and from Table 19 we have the following table:—

TABLE 20

	D.F.	S.S.	M.S.	V.R.
x_1 only ..	1	2·6690		
due to x_2 ..	1	0·0464	0·0464	0·5
Residual ..	2	0·1846	0·0923	
Total	4	2·9000		

This table shows that nothing has been gained by bringing in the number of stems at harvest (x_2) in that, for a given number of shoots in March, the yield of those shoots is not significantly associated with the number of stems they produced. As a matter of fact, this last variance ratio merely measures the significance of b_2, so that, if the data are arranged in this way, there is no need to go through the calculation made above for the determination of the ratio of b_2 to its standard error. We saw that t was 0·7, and if the calculation be carried to an extra decimal place or two it will be seen that this is just the square root of the above variance ratio of 0·5.

A similar table can determine for us the significance of the difference between b_1 and b_2 without determining the separate variances and covariances. If we work out from Table 17

$$\frac{\{(x_1y) + (x_2y)\}^2}{(x_1^2) + (x_2^2) + 2\,(x_1x_2)}$$

we get 2·5708. This is the part of $(y^2) = 2{\cdot}9$ which is "due to" $x_1 + x_2$, and the remainder, namely 0·3292, represents the residual, with 3 D.F. But we know from Table 19 that the 2 D.F. residual is 0·1846. The difference, 0·1446, is "due to" $x_1 - x_2$. This is shown in the following table:—

TABLE 21

	D.F.	S.S.	M.S.	V.R.
$x_1 + x_2$	1	2·5708		
$x_1 - x_2$	1	0·1446	0·1446	1·6
Residual	2	0·1846	0·0923	
Total ..	4	2·9000		

This variance ratio, with 1 and 2 D.F., is not significant, and we note that it is the square of $t = 1{\cdot}3$ obtained as the ratio of $b_1 - b_2$ to its standard error (more exact values are 1·251 for t and 1·566 for V.R.).

The calculations illustrated in Table 20 are often found to be a convenient way of tackling the whole problem from the beginning, instead of proceeding as we did following Table 17. All that need be remembered is how to correct a S.S. or S.P. for regression on a single affecting variate. We have already learnt how to correct a S.S.; a similar correction, illustrated below, applies to a S.P. We start from the data provided by Table 17 and proceed as in Table 22:—

TABLE 22

(x_1^2) = 1642	$(x_1x_2)^2/(x_1^2)$	= 738·2467				
(x_1x_2) = 1101	$(x_1x_2)\,(x_1y)/(x_1^2)$	= 44·3887			D.F.	
(x_1y) = 66·2	$(x_1y)^2/x_1^2)$	=	2·6690	1		(*a*)
(x_2^2) = 786	subtract 738·2467, giving		47·7533			
(x_2y) = 42·9	,, 44·3887, ,,		−1·4887			
(y^2) = 2·90	,, 2·6690, ,,		0·2310	3		(*b*)
	$(-1{\cdot}4887)^2/47{\cdot}7533$	=	0·0464	1		(*c*)
		Difference	0·1846	2		(*d*)

The calculations in the centre are made from the numbers on the left, which are quoted from Table 17, and are self-explanatory. As regards the results on the extreme right, (*a*) is the S.S. due to regression on x_1 only; (*b*) is the residual S.S. of y after eliminating x_1; (*c*) is the S.S. due to

regression on x_2, after x_1 has been eliminated; while (d) is $(b) - (c)$ and is the final residual S.S. $(a) + (c)$ gives the S.S. due to regression on x_1 and x_2, as in Table 19. Note that the S.S. and S.P. have been ordered so that those involving x_1 come in a group at the top, followed by those involving x_2 and y. From this table we can test the significance of regression as a whole, and also that of b_2. b_1 may then be calculated, if needed, by working out $-42{\cdot}9 \times 1{\cdot}4887$ 47·7533, and subtracting this and also 0·1846 from 2·90, finally dividing by 66·2. We may alternatively repeat the calculations with the data re-arranged to show (x_2^2), (x_1x_2) and (x_2y) at the top, and (x_1^2), (x_1y) and (y^2) below, which enables us readily, not only to calculate b_1, but also to test its significance. For details see the schematic Table 24.

The method of Table 22 can be extended indefinitely when there are more than two affecting variables, eliminating first x_1, then x_2 in addition, then x_3 and so on. There is an obvious advantage is so doing, because if at any stage the partial regression coefficient due to the added variate is not significant, it is open to us to reject it, discard the calculations involving the variate and substitute another. Usually there will be no difficulty in choosing the variate with which to start the chain of operations. A general scheme which covers the calculations of both regression and correlation coefficients by this method is given later in Table 24, for the case of two affecting variates, and in Table 30 for three.

Correlation with Two Affecting Variates

We may consider the whole problem afresh from the standpoint of correlation. Thus, in the data of Table 17, x_1 was the number of bean shoots in March. In the experiment it had been intended to sow equal numbers of viable seeds on the plots, which were allocated to different kinds of autumn-sown beans. It was found impossible to do this on a field scale, and at harvest the yields of the bean varieties were observed to be related to the various seed rates. One of the things to be looked for in the harvest results was a possible correlation between yield of beans (y) and number of stems (x_2). For these variates r is, from Table 17

$$\frac{42{\cdot}9}{\sqrt{(786 \times 2{\cdot}9)}} = 0{\cdot}90^*$$

positive and significant at the 5 *per cent.* point. But this is unlikely to be a true measure of the relationship, because both y and x_2 may be affected by the uneven germination, possibly resulting in a spurious correlation. It was impossible to determine the actual number of beans sown on each plot, but plant counts made when the beans were established in January can give a fair measure of the sowing rate. These figures have not been quoted, but we shall assume that a similar measure is given by the number of shoots in March (x_1). It can, in fact, be seen from Table 17 that both y and x_2 follow the trend of the x_1 figures. Let us, therefore, "correct" the components of r for the values of x_1, as was done in Table 22. The numerator (x_2y) has a component $(x_1y)\,(x_1x_2)/(x_1^2)$ "due to" x_1, with 1 D.F., the rest being a residual S.P. with 3 D.F. In the denominator (x_2^2) has a component $(x_1x_2)^2/(x_1^2)$ "due to" x_1, and (y^2) has a corresponding component $(x_1y)^2/(x_1^2)$. We may re-write the calculations of Table 22 in the following form:

Table 23

	D.F.	Regression on x_1	D.F.	Residual	D.F.	Total
(x_2y)	1	44·3887	3	−1·4887	4	42·9
(x_2^2)	1	738·2467	3	47·7533	4	786
(y^2)	1	2·6690	3	0·2310	4	2·9

Now let us work out a correlation coefficient from the residuals, having eliminated the effect of x_1. This value, which we shall call r_2, to denote that it refers to y and x_2, is

$$-\frac{1{\cdot}4887}{\sqrt{(47{\cdot}7533 \times 0{\cdot}2310)}} = -0{\cdot}45$$

This value is tested by comparison with Fisher and Yates, Table VI with 2 D.F., i.e. 1 less than the D.F. of the residuals, just as the r of 0·90 is tested with 3 D.F., 1 less than the D.F. of the S.S. and S.P. in Table 17. The 5 *per cent.* point is − 0·05, and we see, therefore, that after correction for x_1, the correlation coefficient between y and x_2, which was originally + 0·90 and

significant, has become negative, although not significantly so, so that we can deduce that any association there appears to be between yield and number of stems at harvest is accounted for by both variates being associated with the number of shoots in March, i.e. is probably brought about by the uneven sowings on the plots.

The coefficient r_2 is called the *partial correlation coefficient* between y and x_2, the variate corrected for, or "eliminated", being x_1. Such a coefficient may generally be treated as if it were a total correlation coefficient determined from a sample one less in size than the original sample. It should now be noted that the test which we have just carried out for testing the significance of r_2 is mathematically identical with that derived in the previous section for b_2, so that if the problem be tackled by the method of the present section there is no need to calculate b_2 and its standard error. Should, however, the regression equation be desired, it is easily possible to calculate b_2 by multiplying r_2 by the positive square root of

$$\frac{(y^2) - (x_1y)^2/(x_1^2)}{(x_2^2) - (x_1x_2)^2/(x_1^2)} = \frac{0{\cdot}2310}{47{\cdot}7533}$$

i.e. by 0·06956 (see Table 23) in the numerical case we are dealing with (to 4 decimal places r_2 is − 0·4482). Equally, of course, if the problem were tackled by the methods of the last section and b_2 determined, then r_2, if wanted, can be obtained from b_2 by the reverse operation to the one just indicated. Another way to calculate r_2, useful only if the relevant simple correlation coefficients have been calculated, is to subtract the product of the r between y and x_1 and the r between x_1 and x_2 from the r between y and x_2, and to divide the result by the geometric mean of $1 - r^2_{yx_1}$ and $1 - r^2_{x_1x_2}$. The formula is

$$r_{32\cdot1} = \frac{r_{32} - r_{31}r_{21}}{\sqrt{\{(1 - r^2_{31})\,(1 - r^2_{21})\}}}$$

where the numbers in the subscripts may be assigned to any of the variates. In the present case, for example, 3 denotes y, 2 denotes x_2 and 1 denotes x_1, and then $r_{32\cdot1}$ is what we earlier called r_2, the dot denoting that x_1 is the variate eliminated.

There is also a partial correlation coefficient r_1 between y and x_1 for x_2 eliminated. It may be calculated on the lines of Table 23 using, however, (x_1y), (x_1^2) and (y^2), and taking out the components "due to" x_2. The value of r_1 is found to be 0·82, somewhat less than the value 0·95 required for significance. The test, in fact, is equivalent to that for b_1 in the last section, while if we require b_1 by the present method we multiply r_1 by the positive square root of

$$\frac{(y^2) - (x_2y)^2/(x_2^2)}{(x_1^2) - (x_1x_2)^2/(x_2^2)}.$$

Equally we may obtain r_1 from b_1 by the reverse process. The above general formula for $r_{32\cdot1}$ may also be used if for 3 we read y, for 2 x_1 and for 1 x_2.

As a single measure of the association between the affected variate y and the affecting variates x_1 and x_2 there is the *multiple correlation coefficient* R, which is a generalization of r as used as the measure of association with only one affecting variate. It is defined as the positive square root of R^2, where R^2 is the fraction of (y^2) which is "due to" the regression. It is therefore most simply determined from Table 19. We have at once

$$R^2 = \frac{2{\cdot}7154}{2{\cdot}9000} = 0{\cdot}936,\ R = 0{\cdot}97$$

with 2 D.F. In general it is $(b^2)/(y^2)$, where (b^2) is defined below Table 19. It may also be read off from Table 22. Unlike r, R is a coefficient which is restricted to lie between 0 (for no association) and 1 (for a perfect association). It is possible to test its significance as r is tested. The appropriate table (Wishart, 1928, *Quart. J. R. Meteor. Soc.*, LIV, 258, or Snedecor, 1946: *Statistical Methods*, Table 13.6), which corresponds to Fisher and Yates, Table VI, shows that for 2 affecting variates and 2 D.F. the 5 *per cent.* point is 0·975. To three decimals our R is 0·968, so that it falls short of significance. But note that this is not a new test. We have already tested for the significance of regression in Table 19, and saw that it was measured by a variance ratio of 14·7 with 2 and 2 D.F. This test is mathematically identical with that for R using the special table referred to above, and so, just as in the case of simple regression with one affecting variate, we

see that there are two alternative ways of testing significance of the same effect. Obviously R need not be calculated at all unless we need a coefficient, limited to lie between 0 and 1, for comparison with a simple r. The method of Table 19 is the better of the two; in the present case, for example, we might be inclined to say, given R, that it was almost significant, or, if we had confined ourselves to 2 decimal places, that it was significant. Table 19, on the other hand, shows that the variance ratio of 14·7 is well short of the 19·0 required for significance.

R^2 may also be expressed as $\{b_1 (x_1 y) + b_2 (x_2 y)\}/(y^2)$, where b_1 and b_2 are the partial regression coefficients defined below Table 17. In terms of the simple correlation coefficients we have

$$1 - R^2_{3\cdot 12} = (1 - r^2_{12} - r^2_{13} - r^2_{23} + 2r_{12}r_{13}r_{23})/(1 - r^2_{12})$$

where subscripts have been added to R^2 to denote that it is the square of the multiple correlation of y on x_1 and x_2. This formula may, incidentally, be used to calculate either of the other two coefficients $R_{1\cdot 23}$ or $R_{2\cdot 31}$, when we see that the only difference on the right-hand side is that the denominator becomes $1 - r^2_{23}$ in the first case and $1 - r^2_{13}$ in the second. This has no relevance to the present problem, where y is clearly the affected or dependent variate and only one R may be said to have any meaning.

Another useful formula is:—

$$1 - R^2_{3\cdot 12} = (1 - r^2_{31})\,(1 - r^2_{32\cdot 1})$$

or

$$(1 - r^2_{32})\,(1 - r^2_{31\cdot 2}).$$

We saw above, for example, that r_{32} was 0·90 and $r_{31\cdot 2}$ was 0·82. Inserting these values in the second of the two forms we have $1 - R^2_{3\cdot 12} = 0{\cdot}062$, whence $R_{3\cdot 12} = 0{\cdot}97$.

A general scheme of computation which does everything that is necessary for calculating both regression and correlation coefficients, and includes all tests of significance, is shown below in Table 24. A shortened notation, which is self-explanatory, is used, and it will be noted that each stage involves a simple computation that can be done in one step on any ordinary calculating machine. We suppose that there are n sets of observations.

TABLE 24

$(x_1^2) = A$					B			
$(x_1x_2) = P$					P			
$(x_1y) = Q$					R			
.....								
$(x_2^2) = B$	$B - P^2/A$	$= B_1$			A	$A - P^2/B$	$= A_2$	
$(x_2y) = R$	$R - PQ/A$	$= R_1$			Q	$Q - PR/B$	$= Q_2$	
.....								
$(y_2) = C$	$C - Q^2/A$	$= C_1$	$C_1 - R_1^2/B_1 = C_{21}$		C	$C - R^2/B$	$= C_2$	

Estimate	*Test of Significance*
Regression as a whole.	Compare M.S. from $C - C_{21}$ and C_{21} (2 and $n - 3$ D.F.).
$R^2_{y\cdot 21} = 1 - C_{21}/C$.	As above, or Snedecor, Table 13.6.
$b_{y1} = Q/A$.	Compare M.S. from $C - C_1$ and C_1 (1 and $n - 2$ D.F.).
$r^2_{y1} = 1 - C_1/C$ (r same sign as Q).	As above, or Fisher and Yates, Table VI (D.F. $= n - 2$).
$b_{y2\cdot 1} = R_1/B_1$	Compare M.S. from $C_1 - C_{21}$ and C_{21} (1 and $n - 3$ D.F.).
$r^2_{y2\cdot 1} = 1 - C_{21}/C_1$ (r same sign as R_1).	As above, or Fisher and Yates, Table VI (D.F. $= n - 3$).
$b_{y1\cdot 2} = (C - C_{21} - RR_1/B_1)/Q$.	
.....	
or	
$b_{y1\cdot 2} = Q_2/A_2$,	
or $b^2_{y1\cdot 2} = (C_2 - C_{21})/A_2$ (b same sign as Q_2).	Compare M.S. from $C_2 - C_{21}$ and C_{21} (1 and $n - 3$ D.F.).
$r^2_{y1\cdot 2} = 1 - C_{21}\ C_2$ (r same sign as Q_2).	As above, or Fisher and Yates, Table VI (D.F. $= n - 3$).

Note.—The calculations to the right of the vertical rule in the top part of the table are only required for the calculations below the horizontal rule in the bottom part, i.e. for $r_{y1\cdot 2}$ or for the significance of $b_{y1\cdot 2}$. Only the sign of Q_2 is really needed, and, of course, $C_{12} = C_{21}$. A calculation like $B - P^2/A$ is most easily carried out by determining $AB - P^2$ and then dividing by A, so that, given a calculating machine, nothing need be written down except the final answer.

Difference of Regression Functions from Two Independent Samples

Corresponding to the problem dealt with in the fifth section, a solution to which was required before the analysis of covariance procedure could be demonstrated, we now turn to the case where we have two independent samples, each having one affected and two affecting variates. For each there will be a measure of regression as a whole for y on x_1 and x_2, and we then require to know whether these measures for the two samples differ significantly from one another. We could, of course, develop a test for the difference between the b_1's of two independent samples, or between the b_2's, but it is more in keeping with the object of this bulletin, which is to describe the analysis of covariance procedure, to use a single measure to indicate the way in which x_1 and x_2 together affect y, and then to compare such measures in different samples. As before, we shall assume that the residual variances are the same in the two populations. It will be convenient for the purposes of the required test to make some use of the letter notation used in Table 24, i.e. $(x_1^2) = A$, $(x_1x_2) = P$, etc., distinguishing the second sample by the use of dashes. Following out the lines of the right-hand side of Table 6 we may construct the following table of analysis of variance in which the parts due to regression (b^2) and to residual (d^2) are worked out for the two samples separately and together. The samples are taken to be of sizes $n_1 + 1$ and $n_2 + 1$.

Table 25

Sample	D.F.	Regression	D.F.	Residual
1	2	(b^2)	$n_1 - 2$	(d^2)
2	2	(b'^2)	$n_2 - 2$	(d'^2)
sum	4	$(b^2) + (b'^2)$	$n_1 + n_2 - 4$	$(d^2) + (d'^2)$
1 + 2	2	$(\bar{b}^2)$		
diff.	2	$(b^2) + (b'^2) - (\bar{b}^2)$		

M.S. for difference of regressions $= \frac{1}{2}\{(b^2) + (b'^2) - (\bar{b}^2)\} = s_1^2$

M.S. (residual) $= \{(d^2) + (d'^2)\}/(n_1 + n_2 - 4) = s_2^2$

V.R. for difference of regressions $= s_1^2/s_2^2$ D.F. $= 2$ and $n_1 + n_2 - 4$.

Explanation.—The full table of S.S. and S.P. has not been written down. It should be clear, however, that (b^2) is just the quantity $C - C_{21}$ of Table 24 (alternatively calculated as for Table 19), and that (d^2) is C_{21}; furthermore, that (b'^2) is the result of a corresponding calculation on the S.S. and S.P. for the second sample, i.e. $C' - C'_{21}$, and (d'^2) is C'_{21}. For the combined samples the same calculations are repeated on $A + A'$, $P + P'$, etc. Here we call the S.S. due to regression $(\bar{b}^2)$ to denote that it is calculated by determining *mean partial regression coefficients* $\bar{b}_1$ and $\bar{b}_2$, where

$$\bar{b}_1 = \frac{(B + B')(Q + Q') - (P + P')(R + R')}{(A + A')(B + B') - (P + P')^2} = \frac{(Q + Q')_2}{(A + A')_2}$$

$$\bar{b}_2 = \frac{(A + A')(R + R') - (P + P')(Q + Q')}{(A + A')(B + B') - (P + P')^2} = \frac{(R + R')_1}{(B + B')_1}$$

and then working out the S.S. due to the regression function made up of these coefficients. The first form corresponds to the formulæ following Table 17, the second to the notation of Table 24, where we must remember to work all the time on the sums of the S.S. and S.P. for the two

samples taken together. In the present case, however, there is no need to calculate $\bar{b}_1$ and $\bar{b}_2$; we merely require $(C - C') - (C - C')_{21}$, exactly as calculated for the samples separately.

The remainder of the calculations in Table 25 are self-explanatory. In the usual way we infer that the regression relationships are significantly different in the two samples if the V.R. exceeds the 5 *per cent.* value given in the Variance Ratio Table (Fisher and Yates Table V) for 2 and $n_1 + n_2 - 4$ D.F.

To show that what is measured is a function of the difference between the regressions, let us write (b^2) as

$$Ab_1^2 + 2Pb_1b_2 + Bb_2^2$$

one of the forms given earlier. There will be a similar expression, with all letters dashed, for (b'^2) in the second sample. For the combined sample, $(\bar{b}^2)$ will be

$$(A + A')\,\bar{b}_1^2 + 2\,(P + P')\,\bar{b}_1\bar{b}_2 + (B + B')\,\bar{b}_2^2.$$

What we have worked out above is $(b^2) + (b'^2) - (\bar{b}^2)$. This may be shown to resolve itself into an expression of the form of (b^2), i.e. in three parts, but with the first part

$$A\,(b_1 - \bar{b}_1)^2 + A'\,(b_1' - \bar{b}_1)^2$$

and the second

$$2P\,(b_1 - \bar{b}_1)\,(b_2 - \bar{b}_2) + 2P'\,(b_1' - \bar{b}_1)\,(b_2' - \bar{b}_2)$$

while the third part is

$$B\,(b_2 - \bar{b}_2)^2 + B'\,(b_2' - \bar{b}_2)^2.$$

The first part is a *weighted* sum of squares of the deviations of b_1 and b_1' from $\bar{b}_1$. The second is a weighted sum of products of the deviations of b_1, b_2 and b_1', b_2' from $\bar{b}_1$, $\bar{b}_2$; while the third part is similar to the first, but involving the deviations of b_2 and b_2' from $\bar{b}_2$. We are therefore measuring, in a composite kind of way, the differences between b_1 and b_1' and between b_2 and b_2'.

Apart, therefore, from the rather more complicated form of (b^2), which involves the working out of all S.S. and S.P., no new principle is involved in the test of significance of the difference in regressions from that involved in the case discussed earlier in this bulletin where there was only a single affecting variate. Not only so, but the test may readily be extended to one of homogeneity of regressions in more than two independent samples. With, say, k samples, we shall build up an expression of k terms of the form $(b^2) + (b'^2) + \ldots$, from which we subtract $(\bar{b}^2)$ calculated from the S.S. and S.P. $A + A' + \ldots$, $P + P' + \ldots$, etc. The result will be a quantity with $2\,(k - 1)$ D.F. composed of three parts as before, but each part will be the sum of k components, of the form $A\,(b_1 - \bar{b}_1)^2$ in the first part, $2P\,(b_1 - \bar{b}_1)\,(b_2 - \bar{b}_2)$ in the second and $B\,(b_2 - \bar{b}_2)^2$ in the third. The analogy with sums of squares and products of deviations (but of a rather more general kind than are customarily met with) will then be evident. Dividing by $2\,(k - 1)$ we obtain the M.S. s_1^2 measuring the homogeneity of regression. There will likewise be a S.S. for residuals of the form $(d^2) + (d'^2) + \ldots$, having $n_1 + n_2 + \ldots - 2k$ D.F., from which the residual M.S. s_2^2 is obtained. The two M.S.'s are then compared by the Variance Ratio test, with $2\,(k - 1)$ and $n_1 + n_2 + \ldots - 2k$ D.F.

Adjustment of the Observations for Values of the Affecting Variates

We saw earlier how the accuracy of a field trial might be improved by bringing into account an affecting variate such as the yield of the same plot in a previous uniformity trial, and then testing for the significance of "treatments" after adjustment for the values of this variate. A similar problem arises when there are two affecting variates. We might, for example, have separate uniformity trial data for two successive years before, as in the case of many perennial crops, the actual experiment started. Or the relationship between y and x might not be linear, in which case it might be desirable to treat x^2 as a new variate and take it into account as well as x. There is also the problem where the object of the analysis is to probe into the way in which two variates affect a third, the data arising from the somewhat more complicated pattern of a field experiment rather than being of the homogeneous character previously dealt with.

In all these cases the method is as for a single affecting variate, the modifications of the analysis being along the lines of the foregoing section. Thus we first work out an analysis of variance and covariance to include all variates, and then concentrate on the two components of this analysis in which we are interested. One of these will be "treatments", or possibly only part of "treatments", and the other will be the "error". The remainder of the analysis will be as in Table 9, but enlarged to deal with both affecting variates on the lines of Table 25. An example will serve to make the procedure clear.

Example

Brady (1934, *J. Agric. Sci.*, XXIV, 209; 1935, *Suppl. J. R. Statist. Soc.*, **2**, 99), in studying factors influencing lodging in cereals, set up a 9×9 Latin square experiment on oats, in which he tested three varieties Glasnevin Sonas (resistant to lodging), Victory II (intermediate) and Sandy (susceptible), each at three spacings 4 in., 6 in. and 8 in. Among other measurements taken on the plants were thickness of sclerenchyma cell wall in the 5th internode (y), and length (x_1) and diameter (x_2) of this internode. y is a measure associated with degree of resistance to lodging which was believed to be relatively independent of soil variation. It was desired, before suggesting y as the criterion for the isolation of lodging-resistant strains of oats, to determine the nature of its relationship to other morphological characters, in particular length and diameter of internode, particularly as the former of these, at least, was shown to be easily varied by soil fertility conditions. The S.S. and S.P. for y, x_1 and x_2 are shown in Table 26, where "treatments", with 8 D.F., is broken up into components for variety (2 D.F.), spacing (2 D.F.) and interaction (4 D.F.).

Table 26

	D.F.	(x_1^2)	(x_1x_2)	(x_1y)	(x_2^2)	(x_2y)	(y^2)
Rows ..	8	51·45	3·249	−7·307	1·5823	−0·1956	5·3791
Cols. ..	8	68·62	6·153	−5·374	2·0739	0·8653	2·9671
Var. ..	2	262·12	−12·387	−76·107	4·5412	3·5635	22·0983
Sp.	2	24·67	−6·669	−5·992	1·9781	1·9482	2·0929
Inter. ..	4	15·60	1·515	0·942	0·2092	0·0219	0·3647
Error ..	56	142·50	9·940	−16·058	11·0886	1·6874	19·4972
Var. + Error	58	404·62	−2·447	−92·165	15·6298	5·2509	41·5955
Sp. + Error	58	167·17	3·271	−22·050	13·0667	3·6356	21·5901
Inter. + Error	60	158·10	11·455	−15·116	11·2978	1·7093	19·8619

For the partial regression coefficients obtained from the Error line of Table 26 we find:—

$$b_1 = -\frac{194{\cdot}8335}{1481{\cdot}3219} = -0{\cdot}1315, \text{S.E.} = 0{\cdot}0484, t = 2{\cdot}71^{**}$$

$$b_2 = \frac{400{\cdot}0710}{1481{\cdot}3219} = 0{\cdot}2701, \text{S.E.} = 0{\cdot}1737, t = 1{\cdot}55$$

To test the significance of regression as a whole we set up Table 27.

Table 27

	D.F.	S.S.	M.S.	V.R.
Regression ..	2	2·5678	1·2839	4·1*
Residual	54	16·9294	0·3135	
Total	56	19·4972		

We find that y is significantly associated with x_1 and x_2 at the 5 *per cent.* level, although by itself b_2 is not significant. We now wish to test for the effects of varieties, spacings and

interaction in the y variate after adjustment for unequal values of x_1 and x_2. This can all be set out in the single Table 28.

TABLE 28

	(y^2)	(b^2)	(d^2)	Diff.	D.F.	M.S.
Var.	22·0983			2·2618	2	1·1309
Sp.	2·0929			0·4802	2	0·2401
Inter.	0·3647			0·7359	4	0·1840
Error	19·4972	2·5678	16·9294		54	0·3135
Var. + Error ..	41·5955	22·4043	19·1912			
Sp. + Error ..	21·5901	4·1805	17·4096			
Inter. + Error	19·8619	2·1966	17·6653			

The same calculation as was made in Table 27 for the Error line has been made in Table 28 for the lines Varieties + Error, Spacings + Error and Interaction + Error, utilizing the figures at the bottom of Table 26. The differences in the values of (d^2) are then set out on the right-hand side in their appropriate lines, after the manner of Table 12. Finally the M.S. are calculated by dividing by the D.F., and we then see that the only significant result is that the three variety means of thickness of sclerenchyma cell wall, when adjusted to correspond to equal lengths and diameters of internode, differ significantly at the 5 *per cent.* level, the Variance Ratio being 3·6, with 2 and 54 D.F. From the first column of figures in Table 28 it may be seen that varieties when unadjusted were significant at the 0·1 *per cent.* point, and that there was even a suggestion of spacing differences, although in this latter case the Variance Ratio was 3·0, just short of the 5 *per cent.* significance level.

Concentrating now on the three variety means, we may calculate the corrected y means, as is done in Table 29.

TABLE 29

	x_1	x_2	δx_1	δx_2	$b_1\delta x_1$	$b_2\delta x_2$	y	$y-b_1\delta x_1-b_2\delta x_2$
Glasnevin Sonas	9·13	5·59	−1·92	0·30	0·25	0·08	5·12	4·79
Victory II ..	10·56	5·01	−0·49	−0·28	0·06	−0·08	4·71	4·73
Sandy	13·45	5·28	2·40	−0·01	−0·31	0·00	3·87	4·18
Mean	11·05	5·29					4·57	4·57

$$b_1 = -0{\cdot}13,\ b_2 = 0{\cdot}27$$

δx_1 = deviation of x_1 from its mean, and similarly for δx_2. The calculation here may be followed from the table; from an unadjusted variety mean, y, we must subtract b_1 times the deviation of the corresponding mean x_1 from its general mean for all varieties, and also b_2 times the corresponding deviation in x_2. It is then seen that, for plants with the same length and diameter of internode, Glasnevin Sonas and Victory II have practically the same thickness of sclerenchyma cell wall (4·79 and 4·73), whereas the figure for Sandy is decidedly lower (4·18). We may illustrate the testing for significance of the difference between any two of these adjusted means by examining Victory II and Sandy, utilizing the formulæ given earlier for the estimated variances and covariance of b_1 and b_2, and thus extending a formula which was given in a previous section for the corresponding problem with only one affecting variate. The unadjusted difference in y is 4·71 − 3·87 = 0·84, and the corresponding differences in x_1 and x_2 are −2·89 and −0·27 respectively. The adjusted y difference, seen from Table 29 to be 0·55, may then be written as

$$0{\cdot}84 + (2{\cdot}89b_1 + 0{\cdot}27b_2).$$

We require the estimated variance of this quantity. That of the first part is 0·3135, the M.S. for error in Table 28, multiplied by 2/27, or 0·0741, because 0·84 is the difference of two

independent means of 27 plots each. To obtain the estimated variance of the second part we must multiply 0·3135 by

$$\frac{(2{\cdot}89^2)\ (11{\cdot}09) - 2\ (2{\cdot}89)\ (0{\cdot}27)\ (9{\cdot}94) + (0{\cdot}27^2)\ (142{\cdot}50)}{(142{\cdot}50)\ (11{\cdot}09) - (9{\cdot}94^2)} = 0{\cdot}0591$$

obtaining values for (x_1^2), (x_1x_2) and (x_2^2) from the error line of Table 26. Since the estimated variance of the adjusted y difference is obtained by adding the estimated variances of its two independent parts, we can work this out as 0·3135 multiplied by

$$0{\cdot}0741 + 0{\cdot}0591 = 0{\cdot}1332.$$

This yields an estimated variance of 0·04176, or a standard deviation, on extracting the square root, of 0·204. The adjusted y difference, 0·55, divided by its estimated standard error, 0·204, shows that $t = 2{\cdot}7^*$ for 54 D.F. (i.e. the error D.F. in Table 28). This value is significant at the 5 *per cent.* level, the actual 5 *per cent.* point being 2·005. We may conclude that after adjustment of the varieties for unequal length and diameter of internode, Victory II has a significantly thicker sclerenchyma cell wall than Sandy. The same is probably true for Glasnevin Sonas, but a fresh standard error calculation would be required to demonstrate the fact. We noted earlier, in a calculation of the same kind where there was only one affecting variate, that the second term was almost negligible, so that the two adjusted means being compared could virtually be treated as if they were independent means. In the present case this is not so; it will be noted that to add 0·0591 to 0·0741 is nearly to double the estimated variance. It might, therefore, lead to unwarrantable deductions to neglect the exact calculation procedure.

The author's conclusion in regard to this experiment was that thickness of sclerenchyma cell wall was too unstable a character to be of much use as an absolute criterion of the lodging resistant ability of a variety. There was a significant negative correlation between thickness and length of internode, a measure easily varied by soil fertility conditions. Even where soil variation was eliminated as much as possible, by use of the Latin square layout, the differences between the variety means for thickness of cell wall were exaggerated by reason of this correlation. The difference between Glasnevin Sonas and Victory II was significant on the crude figures, but was reduced to nearly zero after adjustment for length and diameter of internode. When grown under sufficiently controlled conditions, easily ascertained measures such as length and diameter of a lower internode would seem to provide as good evidence of the lodging resistant ability of a variety as more difficult measures made on the internal anatomy of the plant. The susceptible character of Sandy is shown clearly by its having a significantly smaller cell thickness than the other varieties even after adjustment to give the estimated cell thickness for plants of this variety having the same length and diameter of internode as the others.

Procedure for More than Two Affecting Variates

The foregoing sections dealing with one affected and two affecting variates have been written up as far as possible to indicate the generality of the method, without using complex algebraic formulæ. To go much further would turn this bulletin into a treatise on multiple regression analysis. But for the reader who wishes to introduce more than two affecting variates a few hints may be given. These chiefly relate to determining the values of the partial regression coefficients b_1, b_2, b_3 etc. from the sample of data. First, we note that the values given earlier for b_1 and b_2 are the solutions of the two simultaneous linear equations in b_1 and b_2

$$\begin{aligned}(x_1^2)b_1 &+ (x_1x_2)b_2 = (x_1y)\\ (x_1x_2)b_1 &+ (x_2^2)b_2 = (x_2y).\end{aligned}$$

When a third affecting variate is present there will be three linear equations of this character,

$$\begin{aligned}(x_1^2)b_1 &+ (x_1x_2)b_2 + (x_1x_3)b_3 = (x_1y)\\ (x_1x_2)b_1 &+ (x_2^2)b_2 + (x_2x_3)b_3 = (x_2y)\\ (x_1x_3)b_1 &+ (x_2x_3)b_2 + (x_3^2)b_3 = (x_3y)\end{aligned}$$

and any one of a number of methods may be used to solve these equations for b_1, b_2 and b_3. The

neatest solution, which at the same time provides the data for calculating the standard errors of the partial regression coefficients, involves inverting the matrix of coefficients of the b's on the left-hand side, a process which is done numerically with the given data. See, for example, R. A. Fisher: *Statistical Methods for Research Workers*, § 29. This gives b_1, b_2 and b_3 as linear functions of the quantities on the right-hand side, and if we use the notation c_{11}, c_{12}, etc., for the coefficients, we have

$$\begin{aligned} b_1 &= c_{11}(x_1y) + c_{12}(x_2y) + c_{13}(x_3y) \\ b_2 &= c_{12}(x_1y) + c_{22}(x_2y) + c_{23}(x_3y) \\ b_3 &= c_{13}(x_1y) + c_{23}(x_2y) + c_{33}(x_3y). \end{aligned}$$

We may now partition (y^2) into a part

$$(b^2) = b_1(x_1y) + b_2(x_2y) + b_3(x_3y)$$

due to the regression, with 3 D.F., and a remaining part

$$(d^2) = (y^2) - (b^2)$$

for deviations from regression, with D.F. 4 less than the size of the sample. The M.S. may be compared to determine the significance, or otherwise, of the regression, and if we denote the M.S. from (d^2) as s^2, we have the estimated standard errors of b_1, b_2 and b_3 given by

$$s\sqrt{c_{11}},\ s\sqrt{c_{22}} \text{ and } s\sqrt{c_{33}}$$

respectively. The remaining analysis should follow without difficulty, since it is chiefly concerned with calculating quantities like (b^2) and (d^2), and the formulæ given in this and earlier sections should suffice, remembering, however, to calculate the D.F. by taking into account the number of affecting variates. It should also be noted that the estimated covariance of b_1 and b_2, say, is s^2c_{12}, and so on for the others. The various coefficients, and tests of significance, may be summed up in Table 30, which is an extension of Table 24 to take account of the additional variate x_3. Note that C, Q and R are differently defined from the former case. This table should enable the reader to go systematically through the computations without having to solve the simultaneous equations as above, and one run through, i.e. up to the horizontal *dotted* line, carries us as far as $b_{y3\cdot21} = b_3$ or $r_{y3\cdot21}$. To obtain a second partial regression or correlation coefficient, e.g. $b_{y2\cdot31} = b_2$, three different calculations are required in the second stage, eliminating x_3 instead of x_2, and these are indicated to the right of the vertical line. The third partial regression coefficient $b_{y1\cdot32} = b_1$ is then obtained by difference. This takes us down to the full horizontal line. It is only if a test of significance is required for $b_{y1\cdot32}$, or if we need to calculate $r_{y1\cdot32}$, that we have to start again at the first stage, eliminating x_2 (or x_3) instead of x_1. This is shown below the horizontal line.

It should be clear from Table 30 that $(b^2) = D - D_{321}$, and that $(d^2) = D_{321}$, which provides the clue to the calculations needed when we wish to compare regression functions from two or more independent samples, or when we wish to apply the analysis of covariance procedure in adjusting the observations (y) for values of three affecting variates (x_1, x_2 and x_3), so far as the complement to Table 28 is concerned. Corresponding to Table 29 we shall have a third component $b_3\delta x_3$ to subtract from y, and the adjusted y difference will then be of the form

$$y' - (pb_1 + qb_2 + rb_3)$$

in which the estimated variance of y' will be that appropriate to the difference between two unadjusted y means, while that of the part in brackets will be

$$p^2\text{var } b_1 + q^2\text{var } b_2 + r^2\text{var } b_3 + 2pq \text{ covar } b_1b_2 + 2pr \text{ covar } b_1b_3 + 2qr \text{ covar } b_2b_3$$

where "var b_1" denotes estimated variance of b_1, and "covar b_1b_2" denotes the estimated covariance of b_1 and b_2, etc. The three estimated variances are equal to s^2 divided by A_{23}, B_{31} and C_{12} respectively (see Table 30), where s^2 is the residual variance obtained by dividing D_{321} by n–4. For the three estimated covariances s^2 is divided by $P_{(123)}$, $Q_{(31)2}$ and $S_{(23)1}$ respectively, and for these we require the extra calculations given at the end of Table 30. Note that the suffices of a letter may be written in any order, e.g. $A_{23} = A_{32}$ and $C_{12} = C_{21}$. Finally the estimated variance of $y' - (pb_1 + qb_2 + rb_3)$ is obtained by summing the estimated variances of its two independent parts, and the t-test follows by dividing $y' - (pb_1 + qb_2 + rb_3)$ by the square root of its estimated variance, yielding a t with n–4 D.F.

TABLE 30

$(x_1^2) = A$					
$(x_1x_2) = P$					
$(x_1x_3) = Q$					
$(x_1y) = R$					
.....					
$(x_2^2) = B$	$B - P^2/A = B_1$				$B_1 - S_1^2/C_1 = B_{31}$
$(x_2x_3) = S$	$S - PQ/A = S_1$				
$(x_2y) = T$	$T - PR/A = T_1$				$T_1 - S_1U_1/C_1 = T_{31}$
.....					
$(x_3^2) = C$	$C - Q^2/A = C_1$	$C_1 - S_1^2/B_1 = C_{21}$			
$(x_3y) = U$	$U - QR/A = U_1$	$U_1 - S_1T_1/B_1 = U_{21}$			
.....					
$(y^2) = D$	$D - R^2/A = D_1$	$D_1 - T_1^2/B_1 = D_{21}$	$D_{21} - U_{21}^2/C_{21} = D_{321}$		$D_1 - U_1^2/C_1 = D_{31}$

Estimate	*Test of Significance*
Regression as a whole.	Compare M.S. from $D - D_{321}$ and D_{321} (3 and $n-4$ D.F.).
$R^2_{y\cdot321} = 1 - D_{321}/D$.	As above, or Snedecor, Table 13.6.
$b_{y1} = R/A$.	Compare M.S. from $D - D_1$ and D_1 (1 and $n-2$ D.F.).
$r^2_{y1} = 1 - D_1/D$ (r same sign as R).	As above, or Fisher and Yates, Table VI (D.F. $= n-2$).
$b_{y2\cdot1} = T_1/B_1$.	Compare M.S. from $D_1 - D_{21}$ and D_{21} (1 and $n-3$ D.F.).
$r^2_{y2\cdot1} = 1 - D_{21}/D_1$ (r same sign as T_1).	As above, or Fisher and Yates, Table VI (D.F. $= n-3$).
$b_{y3\cdot21} = U_{21}/C_{21}$.	Compare M.S. from $D_{21} - D_{321}$ and D_{321} (1 and $n-4$ D.F.).
$r^2_{y3\cdot21} = 1 - D_{321}/D_{21}$ (r same sign as U_{21}).	As above, or Fisher and Yates, Table VI (D.F. $= n-4$).
.....	.....
$b_{y2\cdot31} = T_{31}/B_{31}$.	Compare M.S. from $D_{31} - D_{321}$ and D_{321} (1 and $n-4$ D.F.).
$r^2_{y2\cdot31} = 1 - D_{321}/D_{31}$ (r same sign as T_{31}).	As above, or Fisher and Yates, Table VI (D.F. $= n-4$).
$b_{y1\cdot32} = (D - D_{321} - TT_{31}/B_{31} - UU_{21}/C_{21})/R$.	

To obtain $r^2_{y1\cdot32}$, and tests of significance for $b_{y1\cdot32}$ and $r_{y1\cdot32}$, calculate:—

$A - P^2/B = A_2$		$A_2 - Q_2^2/C_2 = A_{32}$	
$Q - PS/B = Q_2$			
$R - PT/B = R_2$		$R_2 - Q_2U_2/C_2 = R_{32}$	
$C - S^2/B = C_2$			
$U - ST/B = U_2$			
$D - T^2/B = D_2$		$D_2 - U_2^2/C_2 = D_{32}$	

$b_{y1\cdot32} = R_{32}/A_{32}$.	Compare M.S. from $D_{32} - D_{321}$ and D_{321} (1 and $n-4$ D.F.).
$r^2_{y1\cdot32} = 1 - D_{321}/D_{32}$ (r same sign as R_{32}).	As above, or Fisher and Yates, Table VI (D.F. $= n-4$).

For the calculation of the estimated covariances of the partial regression coefficients we require:—

$A - Q^2/C = A_3$		$P_3 - A_3B_3/P_3 = P_{(12)3}$	
$B - S^2/C = B_3$		$Q_2 - A_2C_2/Q_2 = Q_{(31)2}$	
$P - QS/C = P_3$		$S_1 - B_1C_1/S_1 = S_{(23)1}$	

Each time we introduce an additional affecting variate we add one cycle to the calculations. For example, including x_4 we shall have a table like Table 30 beginning (on the left) with five sums of squares and products involving x_1, four involving x_2, three involving x_3, two involving x_4 and finally one for (y^2). The calculations then proceed systematically as for Table 30, but to one more stage until we come to D_{4321}, which completes the calculations necessary for the various tests of significance unless we need other partial regression coefficients besides $b_{y4\cdot321}$. These are obtained by supplementary calculations on the lines of those in Table 30.

Finally, it is worth repeating that when we have reason to suspect that a variate y is related non-linearly to an affecting variate x, the relation may be tested by the same general procedure as has been described, taking x as x_1, x^2 as x_2 etc., and following out the procedure indicated by Tables 24 and 30, or their extension to still more variables. Other functions besides the simple powers of x may be used instead, and, of course, there is nothing to stop us from introducing additional affecting variates at the same time. Thus the quadratic relation between y and x may be examined by having x_1 as x, and x_2 as x^2, while at the same time x_3, x_4, etc., may be introduced as affecting variables to be examined for their linear effect on y. The limit to this procedure is the complexity introduced into the calculations by having quite a number of affecting variates to work with. Modern methods of using special calculating machines of large capacity, e.g. punched-card or electronic, reduce such calculations to manageable dimensions.

PRINTED BY W. HEFFER & SONS LTD., CAMBRIDGE, ENGLAND